OUTLIER CLASSIFICATION CRITERION FOR MULTIVARIATE CYBER ANOMALY DETECTION

THESIS

Alexander M. Trigo, First Lieutenant, USAF

AFIT-ENS-MS-18-M-166

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-18-M-166

OUTLIER CLASSIFICATION CRITERION FOR MULTIVARIATE CYBER ANOMALY DETECTION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Alexander M. Trigo, BS

First Lieutenant, USAF

March 2018

AFIT-ENS-MS-18-M-166

OUTLIER CLASSIFICATION CRITERION FOR MULTIVARIATE CYBER ANOMALY DETECTION

Alexander M. Trigo, BS
First Lieutenant, USAF

Committee Membership:

Dr. Bradley Boehmke
Chair

Dr. Kenneth Bauer
Member

Major Jason Freels
Member

AFIT-ENS-MS-18-M-166

Abstract

Every day, intrusion detection systems catalogue millions of unsupervised data entries. This represents a "big data" problem for research sponsors within the Department of Defense. In a first response to this issue, raw data capture was transformed into usable vectors and an array of multivariate techniques implemented to detect potential outliers. This research expands and refines these techniques by implementing a Chi-Square Q-Q plot-based classification criteria for outlier detection. This methodology has been implemented into an R-based programming solution that allows for a refined and semi-automated user experience for intelligence analysts. Moreover, two case analyses are performed that illustrate how this methodology explicitly identifies outlier observations and provides formal multivariate normality testing to assess the reliability of the techniques being utilized.

Acknowledgments

I would like to thank my family and friends for their reassurance and support throughout my completion of AFIT. Very special thanks are also due to my wife for being extremely patient, understanding, and supportive of my thesis work. Finally, I would like to thank all the faculty and staff who helped me see this project through to completion.

Table of Contents

Page

Abstract iv

Acknowledgments............v

Table of Contents vi

List of Figures ix

List of Tables xi

I. Introduction 1

1.1 Motivation 1

II. Literature Review............3

2.1 Chapter Overview............3

2.2 Cyber Security in the Modern Age............3

2.3 Anomaly Detection Basis 6

2.3.1 Mahalanobis Distance 7

2.3.2 Histogram Matrix............ 8

2.3.3 Factor Analysis 12

2.4 Anomaly Classification 14

2.4.1 Chi-Square Q-Q Plot............ 14

2.4.2 Standard Error of the Estimate. 15

2.5 Multivariate Normality Testing............16

2.5.1 MVN Testing: Mardia's Multivariate Normality Test............ 17

2.5.2 MVN Testing: Henze-Zirkler's MVN Test 18

III. Methodology 19

3.1 Chapter Overview 19

3.2 Implementing an Iterative Chi-Square Q-Q plot 19

3.3 Introducing New Data: Preparation and Cleaning 20

3.3.1 Time Range Observation 23

3.4 Updated Iterative Chi-Square Q-Q Plot 24

3.5 Factor Analysis 25

3.6 Multivariate Normality Testing. 27

IV. Results and Analysis 29

4.1 Chapter Overview 29

4.2 Outlier Classification via Chi-Square Q-Q Plot 29

4.2.1 Outlier Classification of Original Dataset 29

4.2.2 Outlier Classification of Updated Dataset 35

4.3 Factor Analysis 39

4.4 Formal Test for Multivariate Normality 46

4.5 Simulation of Multivariate Normality 49

4.6 Updated Histogram Matrix 52

V. Conclusion 54

5.1 Take-Away 54

5.1.1 Contributions.. 55

5.1.2 Data Considerations 56

5.2 Future Research Considerations 57

VI. Deliverable..58

Appendix A: Rotated Lambda Factor Loadings 59

VII. Bibliography...60

List of Figures

Page

Figure 1: Basic Cyber Network…………………………………………………………...4

Figure 2: Sponsor Data Collection Hierarchy……………………………………………4

Figure 3: Histogram Matrix of a Mail Server Message Distribution…………………..…..9

Figure 4: Cyber-Anomaly Histogram Plot…………………………...…………………..10

Figure 5: Anatomy of a Factor…………………………………………………………...12

Figure 6: Initial Chi-Square Q-Q Plot (Original Data Set)…………………………....30

Figure 7: Reduced Chi-Square Plot with Threshold of .03 (Original Data Set)………...31

Figure 8: Reduced Chi-Square Plot with Threshold of .06 (Original Data Set)……..….32

Figure 9: Error Per Iteration (Original Data Set)……………………………………..33

Figure 10: Initial Chi-Square Q-Q Plot (Updated Data Set)……………………………..36

Figure 11: Reduced Chi-Square Plot with Threshold of .03 (Updated Data Set)………..37

Figure 12: Error Per Iteration (Updated Data Set)……………………………………....38

Figure 13: Horn's Curve vs Sorted Eigenvectors………………………………………39

Figure 14: First Four Factors Based on Rotated Loadings Matrix……………………....41

Figure 15: 2D Scatter Plot of Rotated Factors (Updated Data Set)…………..………….42

Figure 16: 3D Scatter Plot of Rotated Factors (Updated Data Set)……………………...43

Figure 17: 2D Scatter Plot of Rotated Factors (Original Data Set)…………………...…44

Figure 18: 3D Scatter Plot of Rotated Factors (Original Data Set)……………………...45

Figure 19: MVN Test Results (Updated Data Set)………………………………………47

Figure 20: MVN Test Results (Original Data Set)………………………………………48

Figure 21: Rescaled Chi-Square Plot (Original Data Set)……………………………….49

Figure 22: Chi Square Plot (Simulated Data Set)………………………………………..50

Figure 23: MVN Test Results (Simulated Data Set)…………………………………….51

Figure 24: Updated Histogram Plot……………………………………………………...52

List of Tables

Page

Table 1: Original Data Features……………………………………………...…………6

Table 2: Updated Data Features……………………………………………………….22

Table 3: Time Range Feature Addition……………………………………………….24

Table 4: Chi-Square Plot Data Frame…………………………………………………25

Table 5: Data Outlier Classifications (Original Data Set)……………………..……...33

OUTLIER CLASSIFICATION CRITERION FOR MULTIVARIATE CYBER ANOMOLY DETECTION

I. Introduction

Project research sponsors are tasked with defending Department of Defense (DoD) networks from invasive internet-based attacks. Currently, there is a reliance on commercial off the shelf (COTS) solutions to defend against cyber-attacks. These firewalls and intrusion detection services provided by retailers such as McAfee® generate logs when activity is observed. These logs represent a typical big data problem as there is an excess of data and no clear directive for how this data should be analyzed. In this collaboration, the greatest value provided to sponsors is in the exploration and development of any analytic tools that help to both manage and understand their data. The multivariate analytic approach proposed in this body of work are to be used on large static multivariate datasets generated from network traffic logs. Focus for this research is on building and implementing an anomaly classification tool; and testing the performance of the tool to accurately classify anomalies using these multivariate datasets. This is done by adding a meaningful classification criterion based on the Chi-Square distribution and multivariate normality assumptions. Since sponsors have provided an updated raw dataset including new features, an analysis on how reliably anomalies for both the original, and the updated data sets can be classified.

1.1 Motivation

Cyber warfare is an ever-growing front the armed forces are engaging in, and the demand for protection against cyber threats is rapidly increasing. There are many noteworthy examples of the damage cyber assaults can inflict. One such example was the OPM data breach which went unnoticed for 100 days and resulted in the loss of sensitive data for over 20 million employees. A more recent cyber-attack was the “wanna cry ransomware” virus that hit worldwide in early 2017. What was most disturbing about this cyber-attack was not in the chaos it caused, but how easily it was stopped once the kill switch was identified. It was not a solution identified by cyber industry leaders as one would think; rather it was an anonymous programmer [23] who found the simple solution serendipitously. Disturbing stories such as these speak volumes as to the necessity of robust cyber threat prevention and detection, as well as the need for more powerful analytical and data management techniques. Sponsors recognize the immediate need for any data science applications that may be able to help derive meaning from large and complex data logs.

II. Literature Review

2.1 Chapter Overview

This chapter provides greater context for the problem at hand. Intrusion detection systems and firewalls generate a large amount of data that is simply not being utilized effectively. The data on its own does not lend much insight, however, leveraging multivariate analysis techniques, it is possible to identify outliers within these massive datasets. This methodology gives a basis for understanding which vector state blocks may be conveying useful information in terms of anomalous behavior and encourage closer analysis of a block.

With a background established, this chapter concludes with discussion of the techniques required to build an outlier classification tool, and a proposal on how to test for multivariate normality.

2.2 Cyber Security in the Modern Age

It is already understood that the main issue being dealt with is in this scenario is securing DoD networks in the face of modern cyber threats. Raw data is generated from logs in the detection system that compromise the first line of defense against malicious activity. It is comprised of the security system and the Intrusion Detection/Intrusion Prevention System (IDS/IPS). In A Guide to Intrusion Detection and Prevention Systems, intrusion detection is defined as “the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies” [22:9]. Often, firewalls and intrusion prevention systems are functionally very similar in that they both actively analyze the packets from incoming traffic and stop certain traffic from reaching its destination based on predefined protocols. Figure 1 shows what this basic system looks like,

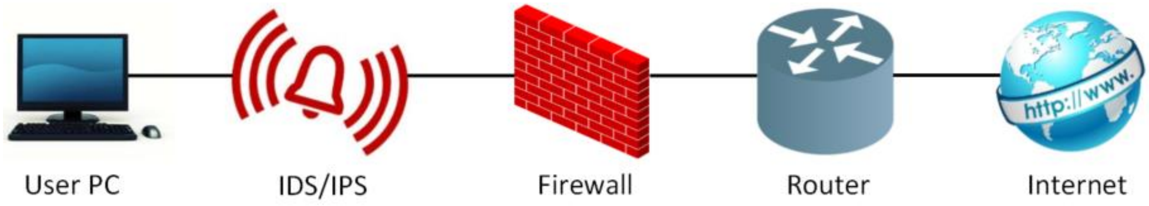


Figure 1: Basic Cyber Network [12]

Whenever an event occurs within the IDS/IPS or the Firewall, a log of that event is created. In the case of the sponsors, many different monitoring systems feed their data into a central data repository which can be accessed via Hadoop. Gutierrez [12] provided a fundamental understanding of how this data collection process occurs.

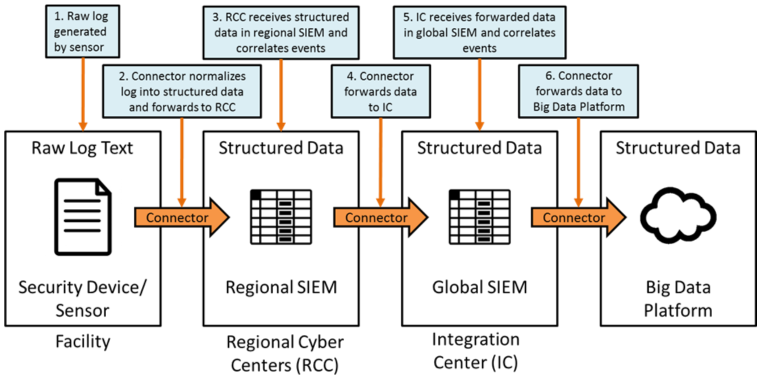
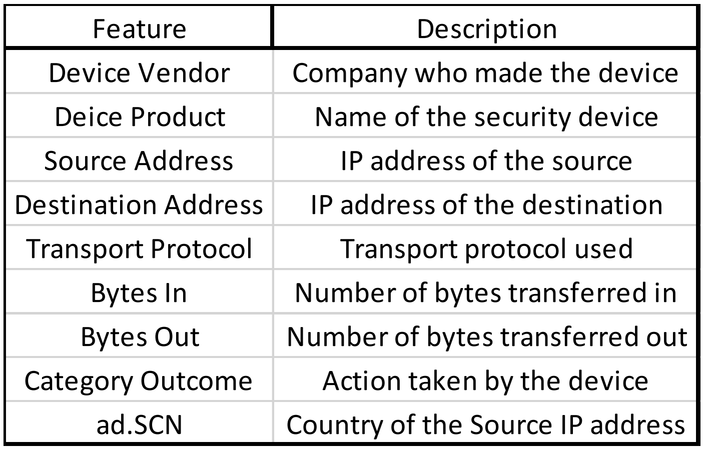

1. Raw log generated by sensor
3. RCC receives structured data in regional SIEM and correlates events
5. IC receives forwarded data in global SIEM and correlates events
2. Connector normalizes log into structured data and forwards to RCC
4. Connector forwards data to IC
6. Connector forwards data to Big Data Platform
Raw Log Text
Security Device/ Sensor
Connector
Structured Data
Regional SIEM
Connector
Structured Data
Global SIEM
Connector
Structured Data
Big Data Platform
Facility
Regional Cyber Centers (RCC)
Integration Center (IC)

Figure 2: Sponsor Data Collection Hierarchy [12]

There are two major problems with this data capture protocol; the first being the sheer quantity of data being collected represents a classic big data problem. Currently, there is more data being collected than there are resources able to analyze it. In a situation such as this, there is always the risk of being overloaded by incoming data [3]. Based on the raw data at hand, it can determine that merely two minutes of data collection from the IDS/IPS systems yield over 30,000 observations. The second issue apparent in this process is that the data is collected in a format that is not ideal for multivariate analysis. Many of the data features of interest tend to be descriptive or categorical rather than continuous. Since this thesis is building upon the body of work done by Gutierrez [12], his solution to these issues will be adopted for continuation of the research. Gutierrez elected to transform the raw log data into tabulated vectors [12]. These tabulated state vectors break and individual feature into multiple features corresponding to its distinct levels. The vector then assigns a block size; and will count how many instances of feature level there are within the span of the defined block size. Blocks partition the original raw dataset into predefined chunk sizes, and these occurrence counters per variable become the feature upon which analysis is conducted. By this method the raw data files may be transformed into a format suitable for multivariate analysis, since categorical variables are effectively turned into continuous counts.

Table 1: Original Data Features



Feature	Description
Device Vendor	Company who made the device
Deice Product	Name of the security device
Source Address	IP address of the source
Destination Address	IP address of the destination
Transport Protocol	Transport protocol used
Bytes In	Number of bytes transferred in
Bytes Out	Number of bytes transferred out
Category Outcome	Action taken by the device
ad.SCN	Country of the Source IP address

Table 1 contains the original features selected from raw IDS/IPS data logs. The feature ‘Category Outcome’ for example is descriptive in nature. There are several different outcomes that may occur. The tabulated state vector will turn each of these outcomes into its own feature and count how many times it occurred within the user defined observational span of raw data (i.e.: How many times category 1 occurs in raw data observations 1-100).

2.3 Anomaly Detection Basis

A desired outcome from this sponsor partnership, is to be able to reliably detect anomalous behavior within the data logs. The term anomaly refers to any observation that varies so far from other observations, that there is a high probability it was generated via alternative means. Identification of anomalies are important because “they indicate significant but rare events and can prompt critical actions to be taken in a wide range of application domains [1:20].” The field of multivariate analysis provides an array of methods that when used together allow the user to effectively deal with large data sets containing multiple features. Before introducing new concepts, several of the multivariate based analytic tools implemented in the original cyber anomaly detection research will be explored. These are important to discuss because they are foundational to all the work conducted in this thesis. Traditionally, multivariate based analytics are very computationally demanding; and have only recently been popularized by the advent of modern computers. These techniques have become invaluable for application in the analysis of large data sets [6].

Once the tabulated state vectors are constructed, two measures are calculated to observe anomalous behavior in the observational level and feature level of the data set. The MD is used to test for any observational outliers within the data set, while the breakdown distances measure feature level departures from the mean. In the original research, these results were plotted simultaneously in a histogram matrix which will be discussed later in further detail. The final core process in this research which carries over from the previous is factor analysis. This powerful technique allows viewing of the group and observations in terms of factors. Factors provide a lot of utility in the ability to understand underlying characteristics of a given data set and allow visualization of patterns within the data.

2.3.1 Mahalanobis Distance

The Mahalanobis Distance (MD) is a measurement technique which determines how far away from the mean a single point in a dataset is [18]. The application of this methodology to the sponsored data set fundamentally drives this strain of research and is what will be expanded upon. The equation for the MD, given by

$$MD = \sqrt{(x-\bar{x})^T C^{-1} (x-\bar{x})}, \tag{1}$$

incorporates the square root of the mean corrected sum of squares multiplies by the inverse of the covariance matrix. The resultant value will indicate how far away from the mean each observation is. Each observation will yield a corresponding score. Higher scores will be associated with observations that are furthest from the data mean, and therefore will be candidates for outlier classification. Since the MD only provides a score at an observational level, the breakdown distances given by

$$BD_i = \left|\frac{(x_i - \bar{x}_i)}{\sqrt{C_{ii}}}\right| \quad (2)$$

are implemented to view how much individual features contribute to anomalous behavior. Once both MD and breakdown distances have been established, they are displayed within a histogram matrix.

2.3.2 Histogram Matrix

When it comes to big data, intuitive and easy to interpret visuals are very important for conveying information efficiently. Achievement of this goal is especially crucial in this line of research, because the individuals the data analysis is intended for will likely not be versed in multivariate applications. In foundational research, the method utilized to convey anomaly information was the histogram matrix.

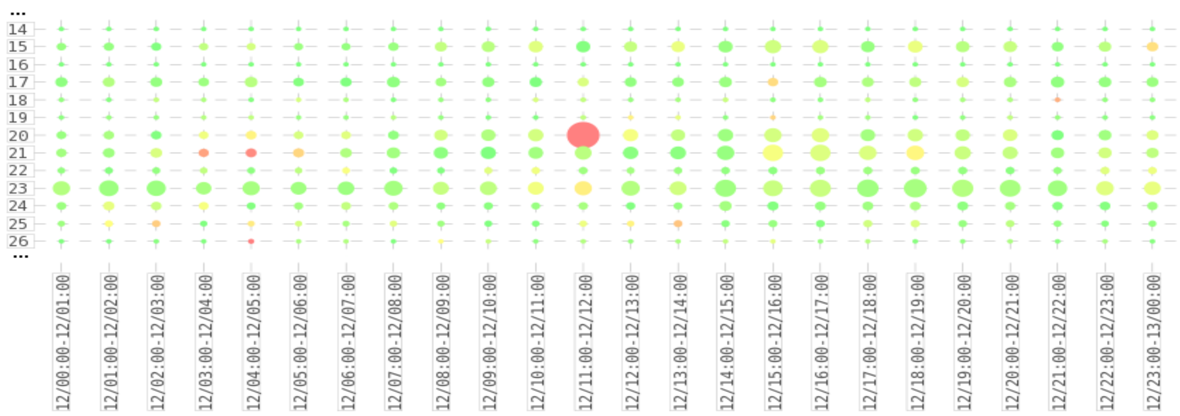


Figure 3: Histogram Matrix of a Mail Server Message Distribution [8]

In figure 3, the size of each dot represents the variable level breakdown distances, while the color of the row conveys the MD. Utilizing some filtering techniques, the intent of this visual is to convey which blocks have corresponding outliers, and which variables most heavily influence outlier classification. Initial implementation yielded results similar to the example matrix shown above and was built into a web-based R Shiny application [4] using a different color scheme.

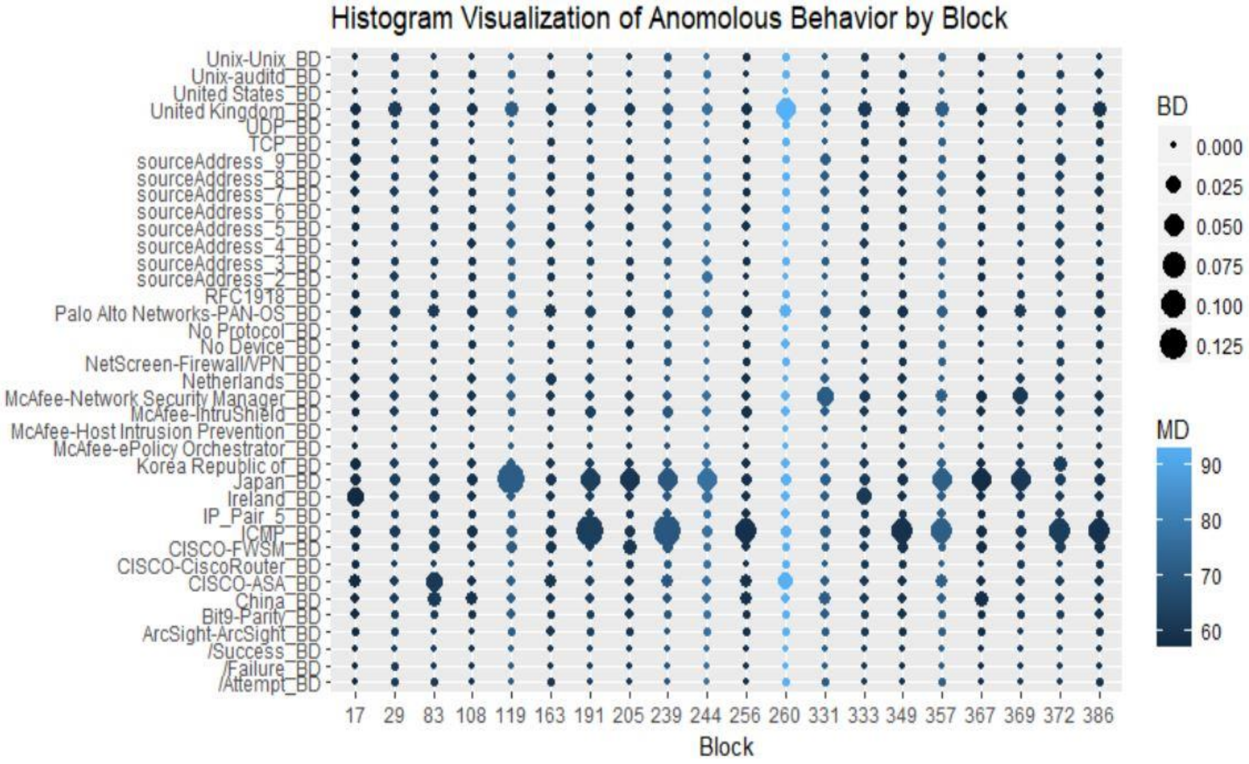


Figure 4: Cyber-Anomaly Histogram Plot

The histogram matrix shown in figure 4 was built in R during this research as an illustrative example. It varies slightly from the one upon which it was designed, however, it captures the same intent. Matrix columns correspond to observation level MD, while the rows correspond breakdown distances. The scales generated indicates to a user that a column portraying an extremely light color is associated with a large MD, while a large dot size corresponds to a large breakdown distance.

Since there exists ambiguity as to which observations should be classified as outliers, a result of this research, is an improvement upon this deliverable. The MD scale is arbitrary, and does not add useful information, especially considering selection criteria. The blocks selected for

display in this histogram were done so simply because they were the observations with the top 20 highest MD scores. There is no rigor in the selection criteria for which observations are plotted in the histogram matrix, rather it is an arbitrary user defined number. Another issue makes itself apparent when trying to determine which block is associated with the most anomalous activity at an observational level. Since MD is represented with a continuous color gradient, and the histogram is plotted from the lowest block to the highest, disregarding the MD, determination becomes difficult. Which block is the highest MD score associated with?
Which block is associated with the second highest MD score? These determinations become difficult and subjective based on this presentation of information. Another problem presents itself when trying to determine where an anomaly takes place. Since the original raw dataset is broken up into equivalently sized chunks, it is unknown exactly which raw observations are associated with the state vector block, and there is no user-friendly way in which to obtain this information. One would have to know how many observations are held within blocks, and manually search for the specified block chunk within the raw data.

The final issue that is not made apparent based on the plot alone lies with the fact that all observations present in this plot along with their corresponding MD scores, are from an unreduced dataset. There is no iterative recalculation of the values seen in this plot upon identification of an outlier. If a point of data is removed, and then all original multivariate values are recalculated, changes in outlier behavior would be expected. While the fundamental concept of this histogram is a good one, there are several key issues that must be addressed. As a product of this research, the histogram plot will be restructured after Chi-Square based classification in order to enhance clarity and saliency of the information conveyed.

2.3.3 Factor Analysis

Another multivariate application discussed in the original anomaly detection work is factor analysis. This technique allows for the identification of underlying “factors” within a data set. Factor analysis asserts that there are hidden factors underpinning a unique data set that give rise to the observed variables, as demonstrated in the following figure:

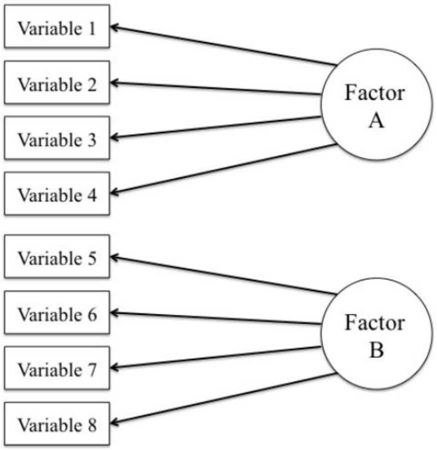


Figure 5: Anatomy of a Factor [5]

In survey-based research, factor analysis is often implemented due to its ability to reveal hidden phenomenon. To illustrate the point, an example of a subject who responds to a battery of questions, giving similar answers to different question categories is salient. If different questions correspond to different variables, such as mental health, family life, job satisfaction, etc., then the potential exists that several variables may describe a hidden factor based on correlations that may not be intuitively realized.

Factors can consist of several or many variables, and often lend themselves to meaningful, intuitive descriptions. The model for factor analysis is given by

$$X = \Lambda f + e \tag{3}$$

in which $XP\ x\ 1$ are all observed responses, $fq\ x\ 1$ are all unobservable common factors, $\Lambda P\ x\ q$ are all factor loadings, and $ep\ x\ 1$ are all unique unobservable factors [2]. The primary concern here lies mainly with the factor loadings matrix given by lambda, as these values allow for the attribution of meaning to the individual factors. It is important not just to find the initial solution of factor loadings, but also at the rotated solution of factors via the rotatefactors() functionality in MATLAB. The decision for which set of factors are to be retained for analysis will be based upon Kaiser's index of factor simplicity [14].

Once factors are determined, further useful information can be gathered by applying factor scores to remaining factors. A factor score is a means in which every observation is weighed to determine its involvement in the factor patterns.

If an observation is more heavily influential in the development of a factor, then it will inherit a higher score [21]. Factor scores given by

$$\hat{f} = X_s R^{-1} \Lambda \tag{4}$$

can be plotted against one another to see if any strong patterns emerge from the underlying data or as another validation step in assessing which data has been classified as being an outlier. For this research, the primary value in factor analysis resides in comparison between factor score plots with results of outlier classification techniques.

2.4 Anomaly Classification

With the background set, this section is dedicated to the exploration of methods that will be implemented in the refinement of Gutierrez's research. Looking beyond the MD of observations, it is important to focus on the properties of the MD in relation to the Chi-Square distribution. Leveraging this relationship with a method for error calculation will allow an implementation criterion by which an outlier may be classified as such. Implementation of the following techniques discussed below will be via the R Studio environment. This initiative will require the functionality of various of R packages available as open source software [4], [24] – [29].

2.4.1 Chi-Square Q-Q Plot

The Chi-Square Quantile-Quantile (Q-Q) plot is a well-established tool for observation of multivariate data structures, and identification of potential outliers. The model is based on a similar study concerned with efficient data analysis of large multivariate data sets generated from geological surveys [7], [9].

The classification of outliers based on the Chi-Square Q-Q plot is contingent upon a property which states that the squared MD calculated for a multivariate normal population is described by a Chi-square distribution. Chi-square values are given by

$$\chi^2(p, r) \qquad (5)$$

$$p = (i - .5)/N \qquad (6)$$

with degrees of freedom, r, equivalent to the number of dataset features, and a probability p [11]. Due to this relationship, squared MD can be sorted in ascending order and plotted against a corresponding set of Chi-Square values. For a perfectly normal multivariate population, a straight line beginning at the origin (0,0) and extending at 45 degrees to some arbitrary distant point such as (50,50) would be observed. This line is an ideal expression of multivariate normality, and often, when a plot of data does not adhere well to this line, it is due to the influence of outliers within the dataset. The Chi-Square Q-Q plot should allow for a visual assessment of data reliability, and obviate any observations associated with outlier activity.

2.4.2 Standard Error of the Estimate.

The standard error of the estimate is a technique implemented often in linear regression. It is used as a measure of accuracy of predictions of a linear model [20]. Described by the following equation, this estimate is the square root of the sum of squares difference between predicted Y' and actual Y observations, divided by the number, N, of observations being considered [17].

$$\sigma_{est} = \sqrt{\frac{\Sigma(Y - Y\prime)^2}{N}} \tag{7}$$

In this outlier detection use case, it is already known that the Y values of an ideal linear model would simply be Chi-square with degrees of freedom and probabilities set as previously described by equations 5 and 6. This is because this is the expression of ideal multivariate normality that is being sought after. While not being derived from a linear regression model, the Y' observations will instead be looked at as the data point given by plotting the calculated MD against the ideal Chi-Square value. As the Standard Error of the Estimate approaches zero, it signals that the multivariate dataset is normally distributed with minimal variance. Obtaining an error estimate of zero would suggest a conformity to multivariate normal distribution with no variability, therefore, minimizing this value is desirable. The value of estimate is not grounded in any intrinsic meaning behind the number itself. It is known that a lower estimate value is desired, and that 0 is an ideal value, but it cannot be determined what values of estimates are acceptable for a dataset. This estimate is not intended to provide an insight into the underlying structure of the dataset, rather, it provides a minimization criterion under which the outlier classification functionality can operate.

2.5 Multivariate Normality Testing

In the field of multivariate analysis, the assumption of multivariate normality underlies many common and parametric analytic techniques. The MD calculation is just one of many techniques in which application requires a multivariate normal dataset with mean μ and covariance matrix Σ. Violation of this assumption can undermine the reliability of any results, and since this work is being conducted for a real-life use case, it is especially important to ensure that this assumption is being addressed. A multivariate normal dataset is one in which there are no multivariate outliers present. When the values of the Chi-square distribution are plotted versus MD squared, there should be a strong visual indication as to how close the dataset is to multivariate normality, however, a formal test for multivariate normality can also be included. The R package ‘MVN’ allows for implementation of unique multivariate normality tests by which can be applied to a given dataset. These tests are a supplement to visual results of the completed Q-Q plot. The results of formal testing should corroborate any initial assessment an analyst might make based on the structure of the plotted data.

2.5.1 MVN Testing: Mardia’s Multivariate Normality Test

The first formal test for multivariate normality proposed by K.V. Mardia [19] builds upon the benefits of univariate normality testing by introducing a means of assessing multivariate measures for skewness ($\hat{\gamma}1, p$) and kurtosis ($\hat{\gamma}2, p$). These measures are used in univariate applications to select a member of a family, such as in the Karl Pearson family, in developing a test for normality, and in investigating robustness of the standard normal theory procedures [16]. For this research, there is a focus on the second application, a function for which is provided by the R based MVN package. The mardiaTest() function calculates the multivariate skewness and kurtosis coefficients, and their corresponding statistical significance p, where .05 is the level of significance. If both the skewness and kurtosis values indicate multivariate normality, then the sample is considered to be multivariate normal based on this test [19]. The authors of the MVN package go on to demonstrate however, that the conclusions of the Mardia test are not always comparable to that of other tests. While it is a commonly accepted measure for normality, it has been criticized that Mardia's skewness measure equals zero not only in the case of multivariate normality, but also within the much wider class of elliptically symmetric distributions [15]. The article offering this criticism advises caution when conducting this test and proposes a new methodology for calculating the skewness measure, however, the test is maintained as it is and a secondary test statistic is employed. The creators of the MVN package offer an example scenario in which different tests yield different conclusions as to the multivariate normality of the given dataset [16]. While this mostly occurs under rare circumstances, where the p value for test statistics is extremely close to the .05 threshold of significance, it is good practice to validate one formal test result via an alternate accepted method.

2.5.2 MVN Testing: Henze-Zirkler's MVN Test

The Henze-Zirkler Multivariate normality test serves the same function as Mardia's test, however, the methodology by which multivariate normality is described is different. This test measures the non-negative functional distance between two distribution functions. It operates on a relationship stating that if the dataset is approximately multivariate normally distributed, then the test statistic returned will in turn follow an approximate log normal distribution. [13]. In the MVN package, a test statistic, HZ, is calculated based on the log normal mean and variance for the dataset at a significance level of .05. The conclusion for multivariate normality is determined based on a test statistic *p* derived from the Hz test statistic. A *p* value lower than .05 would indicate that the dataset is not exhibiting multivariate normal behavior.

III. Methodology

3.1 Chapter Overview

This chapter explores the technical and analytical solutions implemented to refine the outlier classification process.

3.2 Implementing an Iterative Chi-Square Q-Q plot

For every iteration in which an outlier is removed, the covariance matrix, MD, and Chi-Square values are all recalculated, and a new standard error of the estimate is generated based on the updated vectors of new values. This means that the observations associated with the highest MD is removed from the tabulated state vector, leaving with a reduced dataset. For this research, an R function, remove(), was created to perform these tasks. If this function is utilized within a standard for loop, it will index the removed observation at every stage of iteration. Realistically, a perfect standard error of the estimate of 0 is not achievable, but there is a desire to decrease the value by as much as possible while striking a compromise with how much data is being removed. If the best error is achieved after the elimination of 25% of the data, then insight into outlier activity is not truly being provided; and have very likely compromised the reliability of the underlying dataset. From a user perspective, less is more when trying to pinpoint suspicious activity. To address this issue, the function operates on a user defined parameter. The parameter is a percentage of the total data set; and represents the threshold at which iterations will cease if a global minimum is not found. As a default, this parameter is set to 3 percent of the data, meaning that only the top 3 percent of data set observations at a maximum will be considered for anomaly classification. It is desirable to find a local minimum before this threshold, otherwise, there may be underlying issues with multivariate normality assumptions. Once outlier classification is established, final reduced Chi-Square Q- Q plot is generated along with a plot depicting error behavior over an iteration cycle of 75% of the data. This later plot allows for a better scoping for how well the error behaves; and allows comparison to a more global minimum error with the resultant minimum local error from the threshold data range.

3.3 Introducing New Data: Preparation and Cleaning

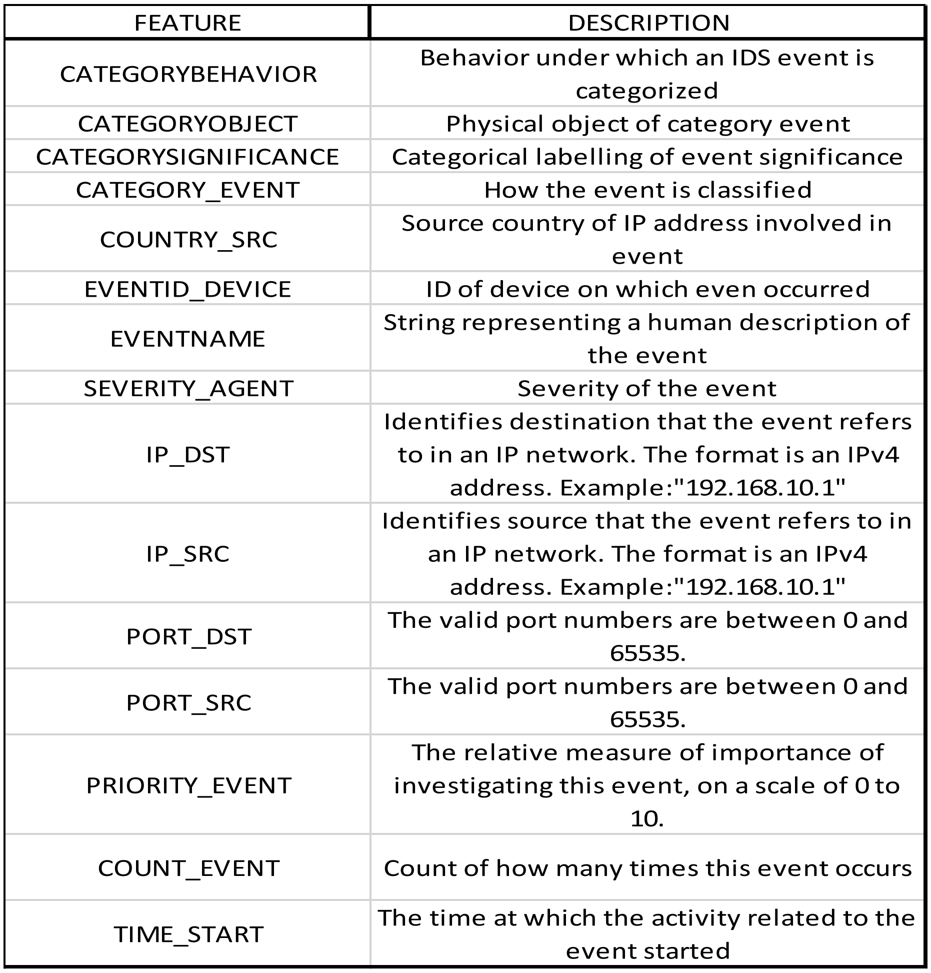
After successful Chi-Square Q-Q plot implementation on the original dataset, the new dataset must be prepared for the same implementation. The way data is prepared is essential to consider when building an automated tool for users to implement easily. The raw data as provided by sponsors contains many features which go unused in the analysis. Functionality must include data preparation before analysis can occur. Changes have been made to which features are being included in the raw data pull. This new data set consists of 93 unique features to be considered for analysis, not all of which are useful. Lacking expertise in cyber security, the best option for determining which features to keep are based on subject matter expert recommendations.

In conversations with research sponsors, it was advised that the analysis be built around the original features, since these are considered the most important from a cyber security perspective. It turns out however, that this simple approach was not possible, since not all of the features available in the original dataset are present in the update.

As many features as possible were carried over from the initial dataset, and several others were chosen based on several criteria. Scarcity of data and levels of data were both used to eliminate features which would likely be useless. For example, a feature missing 90% of the observations, or features which consisted of only a few levels are eliminated from consideration. Several more features were eliminated based on the context in which they were generated. The excel file of raw data is generated in the Hadoop database environment, and several of the features included in the raw dataset originate not from intrusion detection software, but from the data pull itself. These features are frivolous in context of this analysis and were eliminated from consideration.

Of the new features of interest, there is a TIME_START feature specific to the updated raw data which is of importance. This feature allows for the presentation of results in terms of a time and date stamp as opposed to a block size which may be composed of a convoluted non-sequential timeline. The ‘TIME_START’ feature will not be considered during analysis but will allow organization of the dataset in a chronological basis. It is possible to speak about the tabulated state vector blocks in terms of time and date rather than an arbitrary time block, which will convey a much more intuitive and user-friendly experience. All the final features which will be considered for analysis aside from the ‘TIME_START’ are displayed in the following figure along with descriptions as provided by the sponsor.

Table 2: Updated Data Features



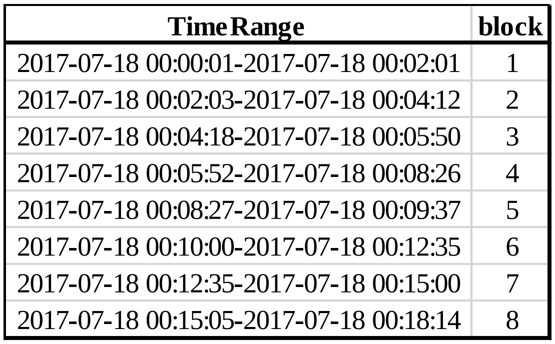
FEATURE	DESCRIPTION
CATEGORYBEHAVIOR	Behavior under which an IDS event is categorized
CATEGORYOBJECT	Physical object of category event
CATEGORYSIGNIFICANCE	Categorical labelling of event significance
CATEGORY_EVENT	How the event is classified
COUNTRY_SRC	Source country of IP address involved in event
EVENTID_DEVICE	ID of device on which even occurred
EVENTNAME	String representing a human description of the event
SEVERITY_AGENT	Severity of the event
IP_DST	Identifies destination that the event refers to in an IP network. The format is an IPv4 address. Example:"192.168.10.1"
IP_SRC	Identifies source that the event refers to in an IP network. The format is an IPv4 address. Example:"192.168.10.1"
PORT_DST	The valid port numbers are between 0 and 65535.
PORT_SRC	The valid port numbers are between 0 and 65535.
PRIORITY_EVENT	The relative measure of importance of investigating this event, on a scale of 0 to 10.
COUNT_EVENT	Count of how many times this event occurs
TIME_START	The time at which the activity related to the event started

To prepare for the technical analysis, these features are extracted from the raw dataset, tabulated in a state vector format, and normalized. This functionality is already available for use via the R based anomalyDetection package resultant from the previous efforts. Sorting the raw data set temporally based on the ‘TIME_START’ variable and setting block size to 50, meaning 50 raw observations per conglomerate group of data, a correlation limit of .9, and a minimum variation of .1, and calling the tabulate state vector function yields a data frame consistent of 1000 observations and 46 variables.

3.3.1 Time Range Observation

Functionality built into the anomaly detection package will yield a vector of observations called blocks. An individual block will correspond to a vectorized and tabulated number of raw observations based on user input, however, these blocks are chronologically ambiguous. Given the new ‘TIME_START’ feature, which follows the POSIXct format, an R function called TimeVector () is built. This function generates an index regarding the current block and truncates a vector of individual time stamps into a single range consistent of the earliest and latest time stamps within that block range. The number of raw observations included within this new truncated time range feature corresponds to the user defined block size to ensure vector dimension consistency between itself and the tabulated state vector. Running the function yields the data frame shown in the following figure.

Table 3: Time Range Feature Addition



Time Range	block
2017-07-18 00:00:01-2017-07-18 00:02:01	1
2017-07-18 00:02:03-2017-07-18 00:04:12	2
2017-07-18 00:04:18-2017-07-18 00:05:50	3
2017-07-18 00:05:52-2017-07-18 00:08:26	4
2017-07-18 00:08:27-2017-07-18 00:09:37	5
2017-07-18 00:10:00-2017-07-18 00:12:35	6
2017-07-18 00:12:35-2017-07-18 00:15:00	7
2017-07-18 00:15:05-2017-07-18 00:18:14	8

Binding a range of times and corresponding block to the tabulated stated vector will give a consistent index and point of reference for the user to determine where an outlier occurred within the raw data. A final output from this research will consist of not only an updated histogram matrix, but a table of information that will make outlier identification much simpler.

3.4 Updated Iterative Chi-Square Q-Q Plot

The procedure for this Chi-Square Q-Q plot development is the same as described in the previous section, except now the updated data set. In contrast to the initial procedure however, the updated dataset contains the 'TIME_START' feature. This means that the TimeVector () function can be ran in order to maintain a time range index associated with outlier classification and subsequent removal throughout the iterations. The MD, corresponding Chi-Square values, block, and time range indexes are bound together in order to prepare for the first iteration.

Table 4: Chi-Square Q-Q Plot Data Frame

MD	block	TimeRange	ChiSqrVal	CumProb
7.617204	889	7/19/2017 3:26	20.79446	0.0005
7.946013	657	7/18/2017 20:34	22.64362	0.0015
9.521463	918	7/19/2017 4:33	23.60629	0.0025
9.653747	690	7/18/2017 21:14	24.2828	0.0035
9.934816	998	7/19/2017 17:04	24.81255	0.0045
9.982936	916	7/19/2017 4:29	25.25186	0.0055
10.0614	910	7/19/2017 4:14	25.62939	0.0065
10.45343	922	7/19/2017 4:45	25.96183	0.0075

Table 4 depicts an excerpt of the initial data frame consistent of 1000 observations. From this point, methodology follows that previously established. A user defined threshold is set, and the function removes observations iteratively until a local minimum for the standard error of the estimate is determined, or until the threshold is reached.

3.5 Factor Analysis

All factor analysis for this research was conducted in the MATLAB programming environment. Up to this point, anomaly detection has been conducted based upon the MD, which is just one multivariate technique. To gain additional perspective on the data, factor analysis-based review of the updated dataset is implemented.

First, a dimensionality assessment of the new dataset, which has already been tabulated and adjusted for multicollinearity is performed. Using simple MATLAB commands, a correlation matrix for the dataset is derived. Subsequent use of the eigs() function returns a matrix of eigenvectors and associated eigenvalues. The dimensionality assessment will be based upon sorting the resultant eigenvalues from highest to lowest and plotting them against horns curve. The number of factors chosen to keep is based upon the intersection of the sorted eigenvalue plot with horns curve. Factors based on the results of the dimensionality assessment are selected since factors with very low eigenvalues do not describe much variability in the dataset.

Before proceeding with analysis of factors, it is determined whether an initial solution or a rotated solution is the most optimal. First, the initial factors are calculated, and a Kaiser score is determined for the given set of factors. This Kaiser sore is compared to the Kaiser score of a rotated set of factors to determine which solution is better. The solution set associated with the higher Kaiser score is used for further assessment. To conclude factor analysis, the dataset features which are associated with the factors are observed. With good data, certain factors tend to correspond to describable phenomenon. It is often a good sign for the dataset when a factor can be described in easily sensible terms. Regardless of the results of factor description, the factor scores for the factors are calculated and then plotted against one another. In a separate color, the data points which were previously classified in the iterative Chi-Square Q-Q plot function are plotted. The primary benefit derived from the factor analysis of the dataset will be in allowing review of classified outliers within the context of an alternative multivariate analytic technique. Factor scatter plots offer an easy visualization of the data and can often be rendered in such a way as to reveal groupings and patterns within data. It is observed whether any patterns of behavior manifest within the factor plots based on grouping from Q-Q plot outlier classification.

3.6 Multivariate Normality Testing.

While the Chi-Square Q-Q plot gives an indication as to the underlying structure of the dataset in terms of multivariate normality, it is done so somewhat subjectively. Before drawing conclusions based on the structure of the Q-Q plot, two formal multivariate normality tests are run. Mardia's Multivariate Normality Test and the Henze-Zirkler Multivariate Normality Test are both conducted to reach consensus as to the underlying nature of the dataset.

To conduct Mardia's MVN test, a multivariate kurtosis and skewness measure of the dataset is calculated. Conveniently, implemention a code-based solution is not required, as this test is already included in the R package `MVN`[16]. Calling the mardiaTest() function with the dataset will yield a table with values for the Mardia test statistic, as well as a p value associated with that statistic. The p value significance for determining multivariate normality is .05 by default, where anything greater than this value is considered multivariate normal. Much like the Mardia function, the ability to evaluate for normality using the Henze-Zirkler MVN Test is built into this package via the hzTest() function. Running the data through this function will yield an Henze-Zirkler test statistic value as well as a corresponding p value used for multivariate normality determination.

These tests are conducted three separate times on the original, reduced dataset, the updated, reduced data set, and finally, on a simulated multivariate normal dataset, whose mean and covariance matrix are derived from the sample mtcars dataset available in R Studio. These formal results allow formation of concrete conclusions about the underlying structure of the data, and to see how well the Chi-Square Q-Q plots were able to visually indicate these results.

Finally, a Q-Q plot and standard error of the estimate value for the multivariate normal dataset is generated to compare differences between it and the two actual datasets.

IV. Results and Analysis

4.1 Chapter Overview

The execution of Chi-Square Q-Q plot functionality for both the original and updated data sets is the first focus of this chapter. The functionality is demonstrated on two separate datasets to highlight differences in results. The results of the Chi-Square Q-Q plots are compared with a factor analysis, and then conclude with formal multivariate normal testing. For purposes of further comparison, the corresponding Chi-Square Q-Q plot, formal multivariate normality test results, and standard error of the estimate are also calculated for a simulated multivariate normal data set.

4.2 Outlier Classification via Chi-Square Q-Q Plot

The following two sections demonstrate the result of outlier classification via the Chi-Square Q-Q plot. The function to generate a Chi-Square Q-Q plot consists of the user defined threshold input, and the standard error of the estimate calculation. Both the original data sets plotted before and after iteration are shown, as well as how the error calculation behaves over many iteration cycles.

4.2.1 Outlier Classification of Original Dataset

For the first implementation of the outlier classification function, the first dataset as seen in table 1 is utilized. All data points corresponding to individual observations are plotted as singular black dots, while the ideal multivariate normal model is plotted as a solid red line.

This initial plot will show two important elements: first, a visual inspection of how closely the data follows the ideal multivariate normal model is visualized, and second, it is observed how outlier activity effects the plot.

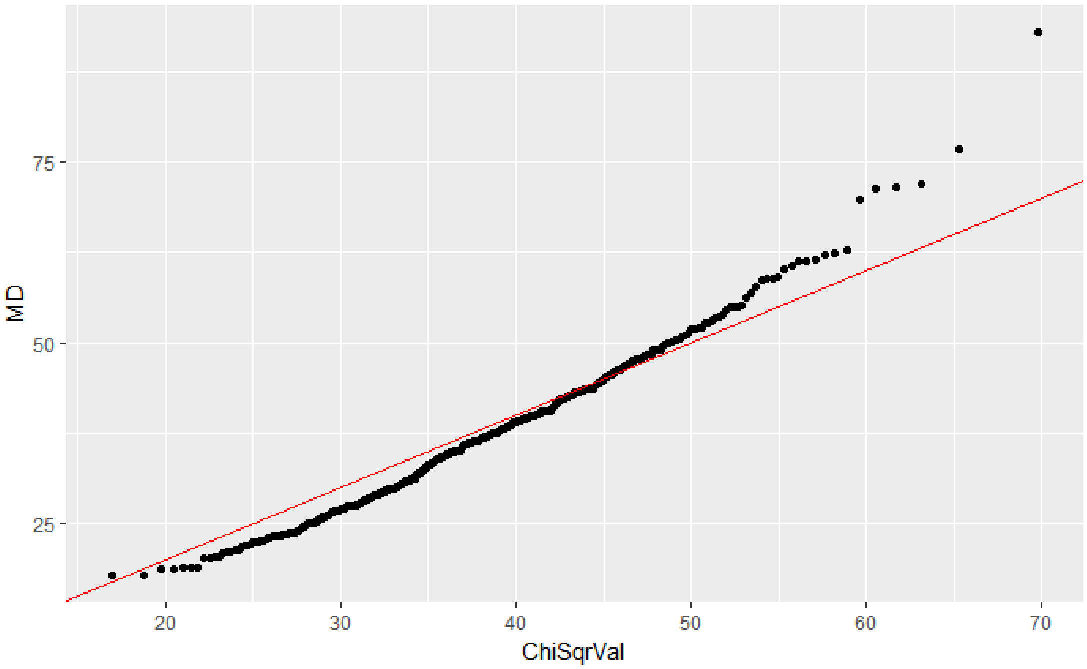


Figure 6: Initial Chi-Square Q-Q Plot (Original Data Set)

Without removing any observations, it is seen that there is a noticeable departure from the multivariate normal model. On visual inspection alone, there are six observations that appear to be consistent with outlier behavior. The core of the data seems to fit the ideal model well, but there is room for improvement. The standard error of the estimate for this plot is 2.79 which the function seeks to minimize within the threshold range. Setting the data keep threshold to the default of 3 percent implies that of the 393 tabulated state vector observations, only the first 11 are considered for outlier classification

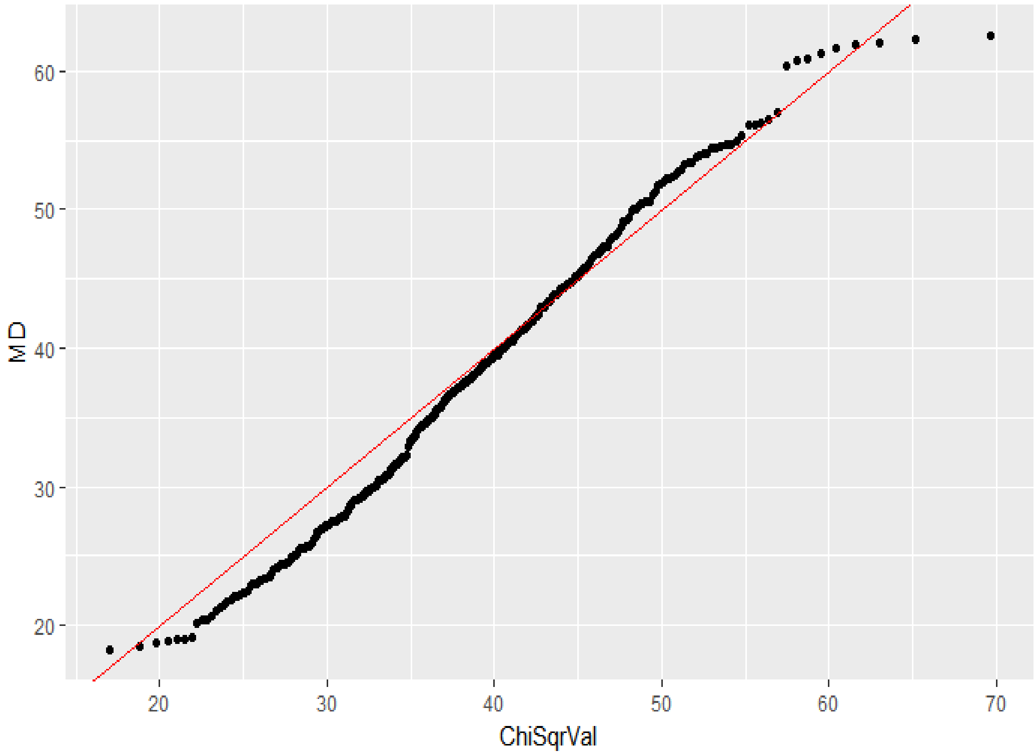


Figure 7: Reduced Chi-Square Q-Q Plot with Threshold of .03 (Original Data Set)

There is a marked improvement in the fit of the data to the ideal model by removing the first 11 observations. Recalculating the Standard Error of the Estimate confirms this observation with an improved estimate of 1.91. Since the function removed observations up to the user defined threshold, this scenario in which the user may wish to expand the threshold to increase the number of observations being considered for anomaly classification.

Although the data is reasonably fit to the ideal model, there are still several points that appear to be deviant. For this reason, the threshold is expanded to 6% and the iterative functionality is implemented to find a new global minimum and replot.

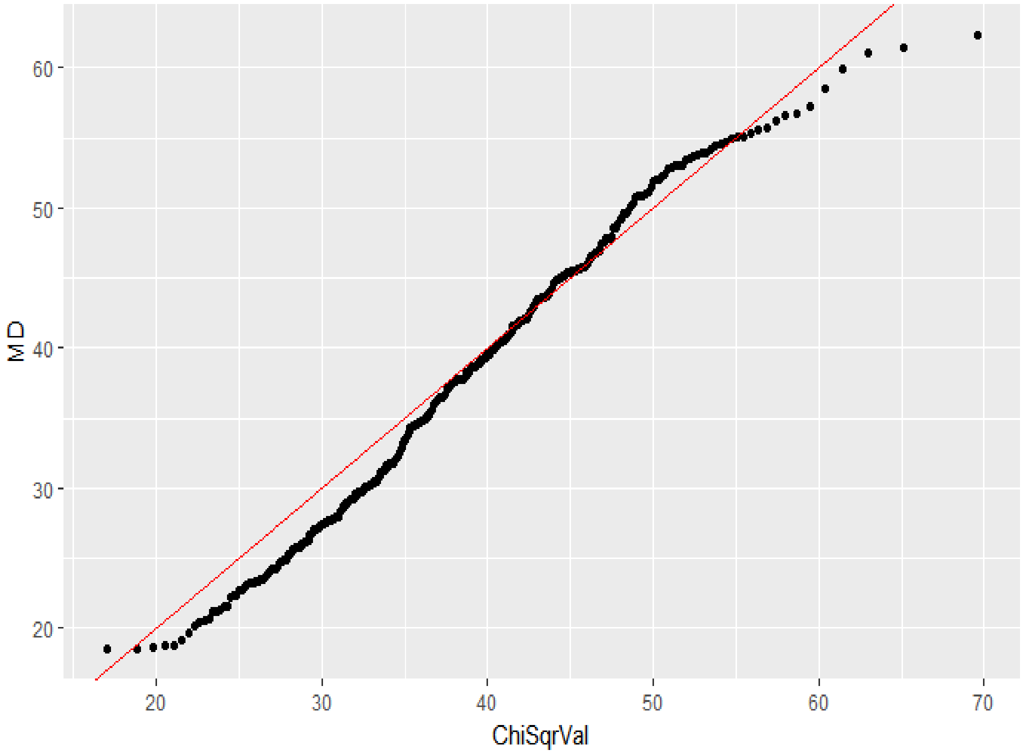


Figure 8: Reduced Chi-Square Q-Q Plot with Threshold of .06 (Original Data Set)

Increasing the threshold means that the initial 23 observations are considered for classification instead of just 11. Despite this higher threshold, the function only requires 16 iterations to find a further improved standard error of the estimate of 1.796. In this case, there is data that appears to fit the ideal multivariate model reasonably well; and yields a local minimum error before the cut-off point. The final plot produced seems to fit the ideal model quite well.

It cannot be concluded with certainty that the remaining dataset is multivariate normal, however, it does look to be a possibility. If it is known that multivariate normality assumptions have been satisfied, then the conclusion can be drawn that the outliers classified were done so correctly. Since this a good model was maintained without eliminating an excessive number of observations, these results seem promising at a glance. To see how error behaves over many iterations, standard error of the estimate for 295 iterations is plotted.

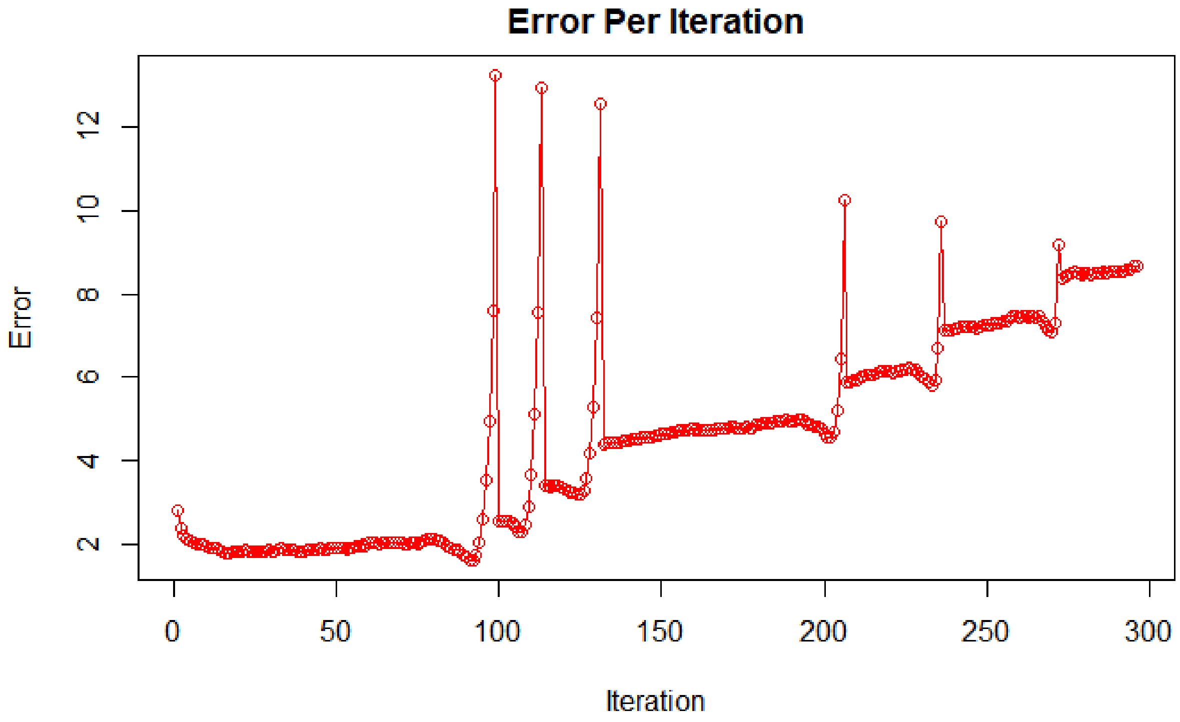
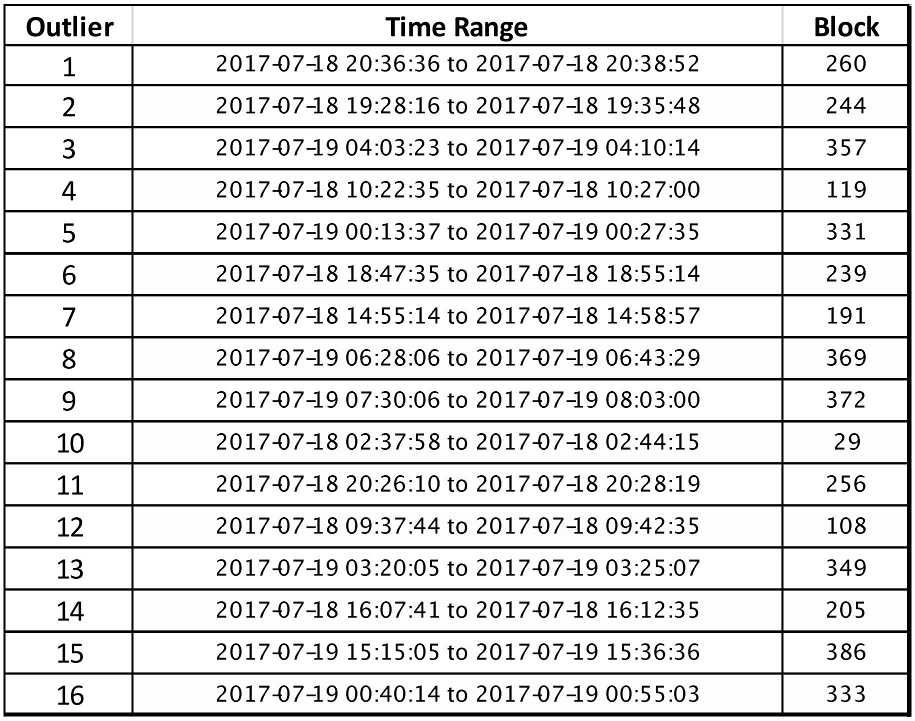


Figure 9: Error Per Iteration (Original Data Set)

A minimum global error of 1.62 is achieved after the 90th iteration, however, this would constitute an excessive elimination of data as well as an extremely large outlier report. A classification of 16 outliers remains optimal, as the detriment of eliminating so much data far outweighs the benefit of marginal error improvement. One benefit from the iterative Chi-Square Q-Q plot function is that outliers are indexed as they are classified, thus making it easy to retrieve once the function has finished executing.

Table 5: Data Outlier Classifications (Original Data Set)



Outlier	Time Range	Block
1	2017-07-18 20:36:36 to 2017-07-18 20:38:52	260
2	2017-07-18 19:28:16 to 2017-07-18 19:35:48	244
3	2017-07-19 04:03:23 to 2017-07-19 04:10:14	357
4	2017-07-18 10:22:35 to 2017-07-18 10:27:00	119
5	2017-07-19 00:13:37 to 2017-07-19 00:27:35	331
6	2017-07-18 18:47:35 to 2017-07-18 18:55:14	239
7	2017-07-18 14:55:14 to 2017-07-18 14:58:57	191
8	2017-07-19 06:28:06 to 2017-07-19 06:43:29	369
9	2017-07-19 07:30:06 to 2017-07-19 08:03:00	372
10	2017-07-18 02:37:58 to 2017-07-18 02:44:15	29
11	2017-07-18 20:26:10 to 2017-07-18 20:28:19	256
12	2017-07-18 09:37:44 to 2017-07-18 09:42:35	108
13	2017-07-19 03:20:05 to 2017-07-19 03:25:07	349
14	2017-07-18 16:07:41 to 2017-07-18 16:12:35	205
15	2017-07-19 15:15:05 to 2017-07-19 15:36:36	386
16	2017-07-19 00:40:14 to 2017-07-19 00:55:03	333

This excel table is a reproduction of the R generated data frame and depicts the block from which each observation was classified, as well as the order in which the blocks were classified. To clarify, in this example, outlier 1 is associated with the first and most anomalous observation that was classified. Additionally, it can be noted that a time range is generated for each block despite the time feature being absent from the original data set.

This feature is merely simulated for demonstration purposes. According to this function, the first outlier occurs in block 260, the second in 244, and so forth. The order in which outliers are recorded matters because it provides a clear rank of severity. This function is operating in an ideal manner for this dataset in which it is possible to minimize a local error with minimal data eliminations. Confidence is also placed in the multivariate methods implemented in outlier detection since there is a strong visual indication of multivariate normality based on the final Chi-Square Q-Q plot. One major area of improvement is in the ability to specificity outlier location. Carrying these methods over, it is possible observe how a completely different dataset performs with this functionality. The only difference in this implementation will be the inclusion of a time vector which is used to keep track of the temporal location of outliers as they are classified.

4.2.2 Outlier Classification of Updated Dataset

To provide a visual basis for the execution of the outlier classification function, the initial representation of the MD vs Chi-Square Q-Q plot is observed.

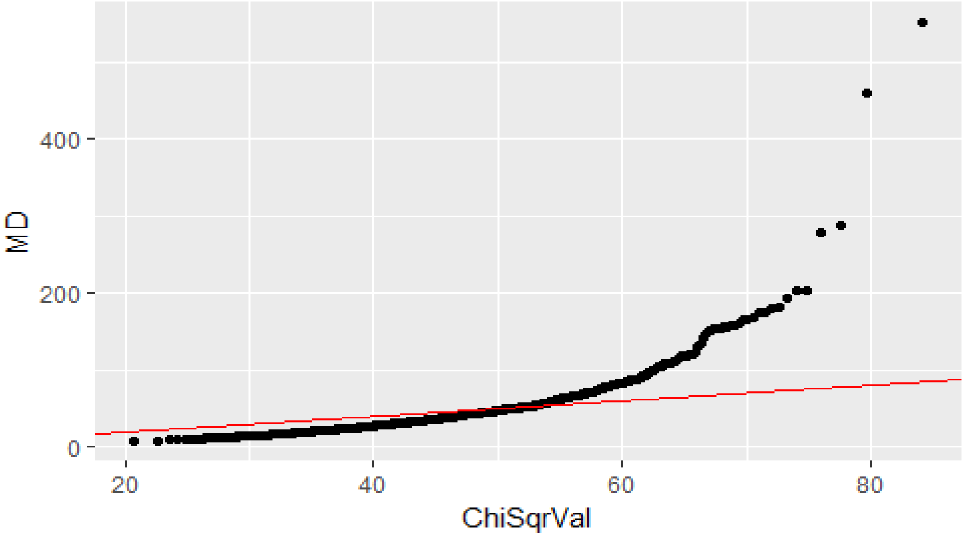


Figure 10: Initial Chi-Square Q-Q Plot (Updated Data Set)

This figure depicts the MD plotted against the corresponding Chi-Square value for all the initial 1000 observations. This plot represents the dataset prior to anomaly classification. The red line plotted along with the points again depicts the model and indicates an ideal multivariate normal model. Just by visually inspecting this graph, it is notable that the first four observations constitute extreme departures from the multivariate normal model. Calculating for the initial standard error of the estimate of this plot yields a very high value of 29.6. There is much room for improvement here, so preparation for iteration is done by setting the threshold to the default of 3 percent, which constitutes 30 observations. In this case, the function finds the global minimum after the 29th iteration. Replotting the graph yields the following updated figure.

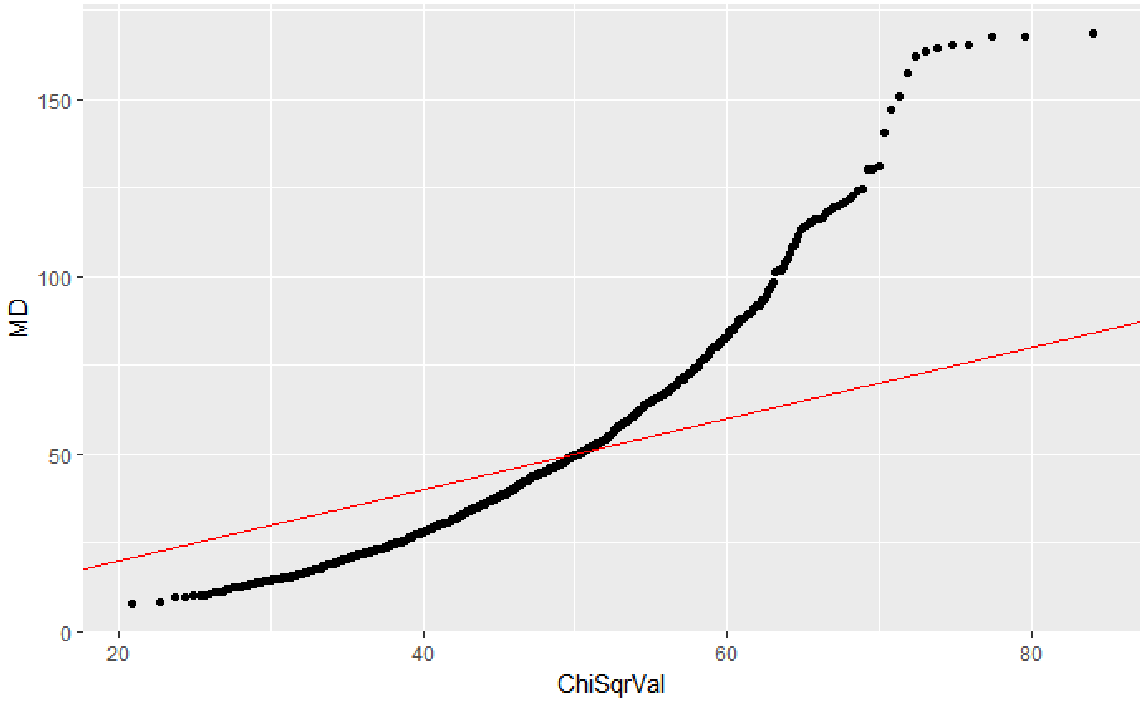


Figure 11: Reduced Chi-Square Q-Q Plot with Threshold of .03 (Updated Data Set)

Despite removing 29 observations, the plot is still exhibiting a severe departure from the model and find an associated error estimate of 17.52. These results do not instill confidence in the reliability of the classification, as the data demonstrates severe departure from a state of multivariate normality. A common temptation is to simply eliminate data until a good error estimation is found, however, it is detrimental to remove an excessive number of observations just to improve error. 750 iterations of this tabulated dataset are made ensuring that with every iteration the covariance and mean of the dataset are recalculated. With every iteration, the highest MD is removed, and the standard error of the estimate saved for that iteration. This enables observation of the behavior of error scores and compare the local minimum found via the threshold confined function with a more global minimum.

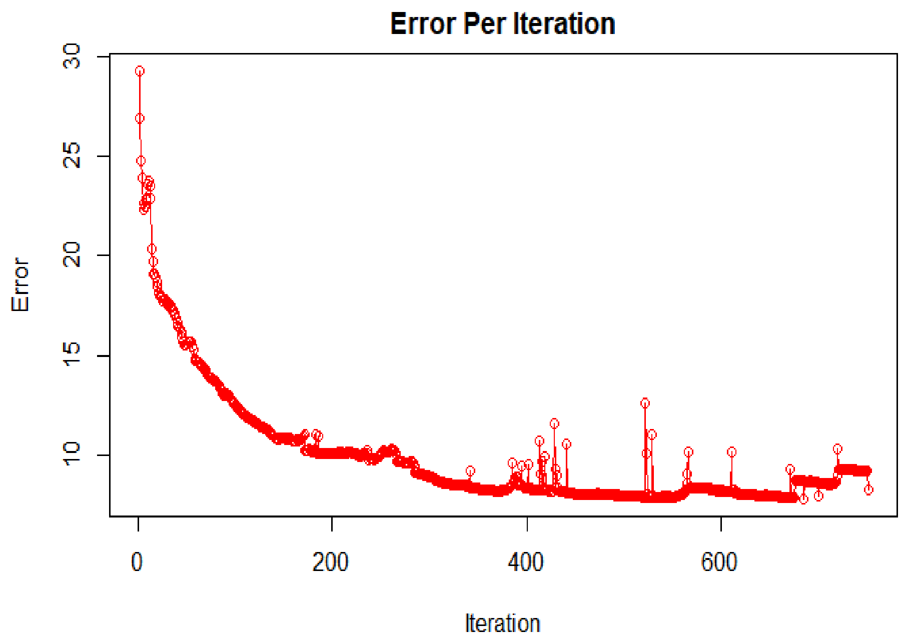


Figure 12: Error Per Iteration (Updated Data Set)

A pronounced drop in error is observed from the starting point to iteration 200 followed by a gradual flattening out at around the 400th iteration. The scale of this graph makes it hard to tell, but It is important to note that despite the error leveling off, the best value schieved from this entire run of iterations is still at 7.82 after the 685th iteration. This value is more than triple that of the initial plot for the original dataset. With these results it does not appear possible to make any reliable recommendations as to which observations should be classified as outliers.

Even if a recommendation is made to look at the first 29 observations classified by the function, that recommendation would be made in light of a Q-Q plot indicating a significant breech of the multivariate normality assumption. This does not mean the classification criteria does not work, it just indicates that the particular dataset used likely contains severe violations of multivariate normality and is not a good candidate for this type of analysis. To validate the concerns regarding this dataset, two separate formal tests for multivariate normality are completed.

4.3 Factor Analysis

Using the updated dataset consistent of all 1000 observations and 46 features, MATLAB functionality is used to find the correlation matrix for the data and the associated eigenvalues and eigenvectors. Sorting eigenvalues from highest to lowest, they are plotted along with Horn's Curve for this unique data set dimensionality.

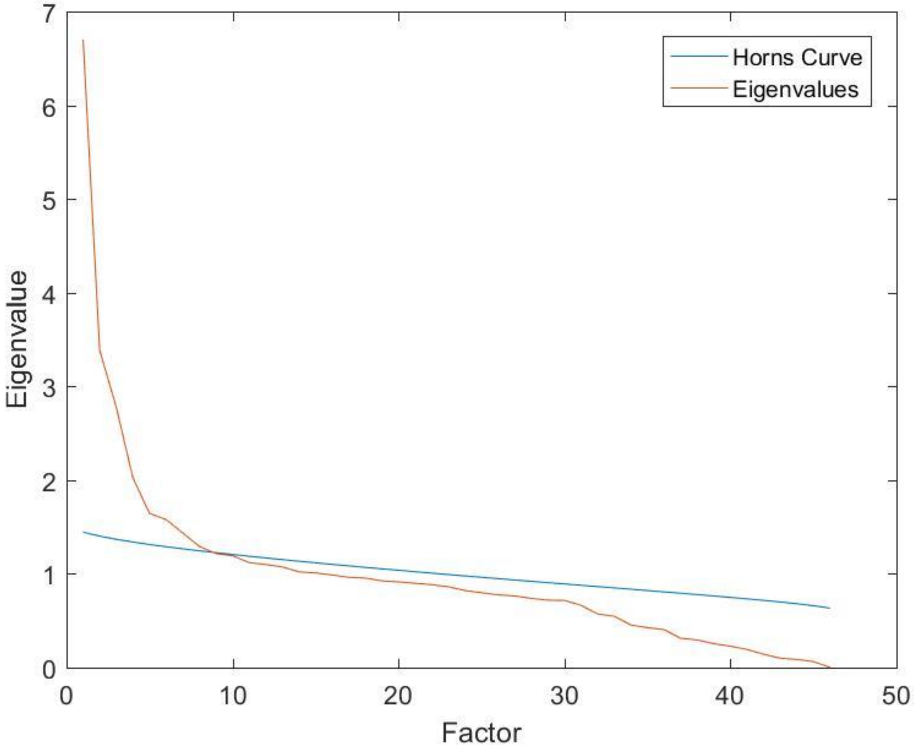
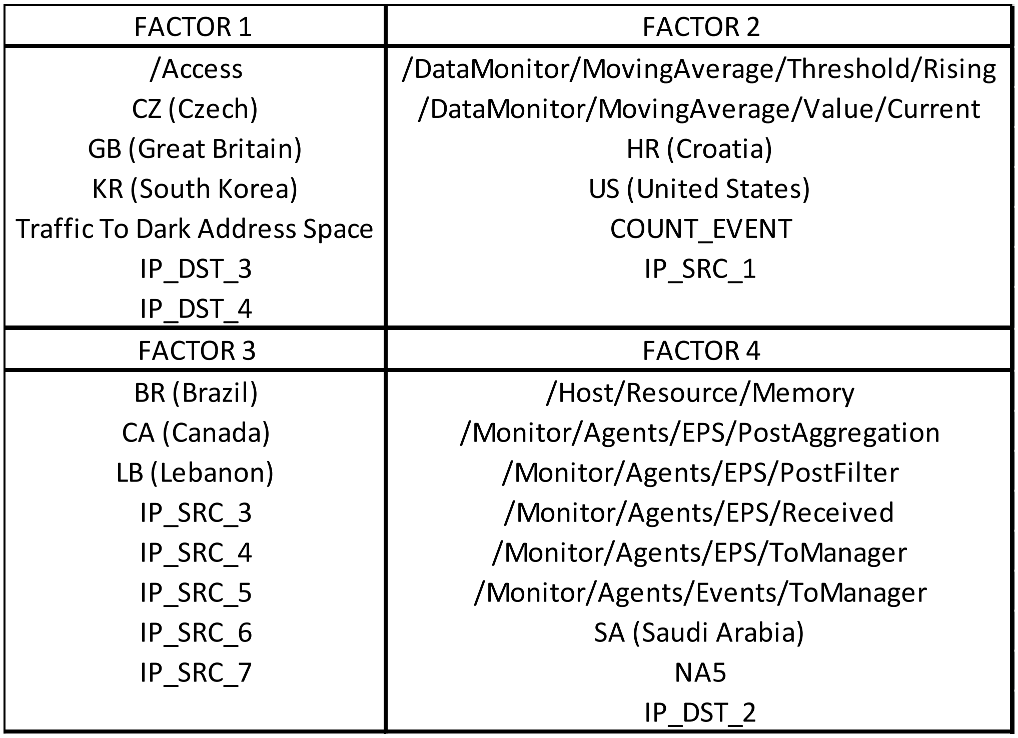


Figure 13: Horn's Curve vs Sorted Eigenvectors

Where the plot of eigenvalues crosses Horn's Curve is often a good cutoff point to justify a dimensionality assessment. Conventional wisdom on dimensionality assessment prescribes the approach of keeping every factor with a corresponding Eigenvalue greater than 1. For this dataset, there are 15 factors which meet this criterion and describe a total of 62% of the data variability. Using the results of the Horn's curve criteria, only the first 10 factors are considered since that is approximately the point at which the eigenvalue plot intersects Horn's curve. Keeping the first ten components, 51% of the variability remains accounted for. In favor of concision, subsequent analysis is based upon the first 10 factors.

Before examining the factor loadings, it must be determined if the initial set of loadings, or the rotated solution will be used. Running the Kaiser score function on both factor sets and find a Kaiser index for initial and rotated solutions of 0.5421 and 0.7735 respectively indicates a better Kaiser score for the rotated solution. The subjective terminology Kaiser ascribes to these values are 'miserable' for the initial factor solution, and 'middling' for the rotated solution. While neither solution set is particularly impressive, an assessment of the rotated solution is made in an attempt to derive meaning based on associated variables. The factor scores of the rotated factor set, from which the following factors are derived, are located in appendix A.



FACTOR 1	FACTOR 2
/Access	/DataMonitor/MovingAverage/Threshold/Rising
CZ (Czech)	/DataMonitor/MovingAverage/Value/Current
GB (Great Britain)	HR (Croatia)
KR (South Korea)	US (United States)
Traffic To Dark Address Space	COUNT_EVENT
IP_DST_3	IP_SRC_1
IP_DST_4	
FACTOR 3	FACTOR 4
BR (Brazil)	/Host/Resource/Memory
CA (Canada)	/Monitor/Agents/EPS/PostAggregation
LB (Lebanon)	/Monitor/Agents/EPS/PostFilter
IP_SRC_3	/Monitor/Agents/EPS/Received
IP_SRC_4	/Monitor/Agents/EPS/ToManager
IP_SRC_5	/Monitor/Agents/Events/ToManager
IP_SRC_6	SA (Saudi Arabia)
IP_SRC_7	NA5
	IP_DST_2

Figure 14: First Four Factors Based on Rotated Loadings Matrix

A strong example of this is the first four factors grouped by their descriptive variables. In factor analysis, the variable groupings within a factor often fit some sort of meaningful description readily, however, in this case, it is difficult to articulate what each factor may correspond to. To a subject matter expert with a more thorough understanding of this particular dataset, these factors potentially reveal a meaningful or interesting phenomenon, however, appropriate labels cannot be properly ascribed for these factors as they appear. One may interpret factor one as being associated with detection of dark web access, or factor 4 as being associated with a monitoring system specifically linked to Saudi Arabia, however, these descriptions are ambiguous and likely falsely represent that which they are intended to clarify.

To conclude the factor analysis, the factor loadings for each of the 1000 observations is calculated. Plotting factor scores for factor 1 against factor scores for factor 2, outlier behavior is sought and to observe this behavior might correspond to the Chi-Square Q-Q plot classified outliers, which are plotted in red.

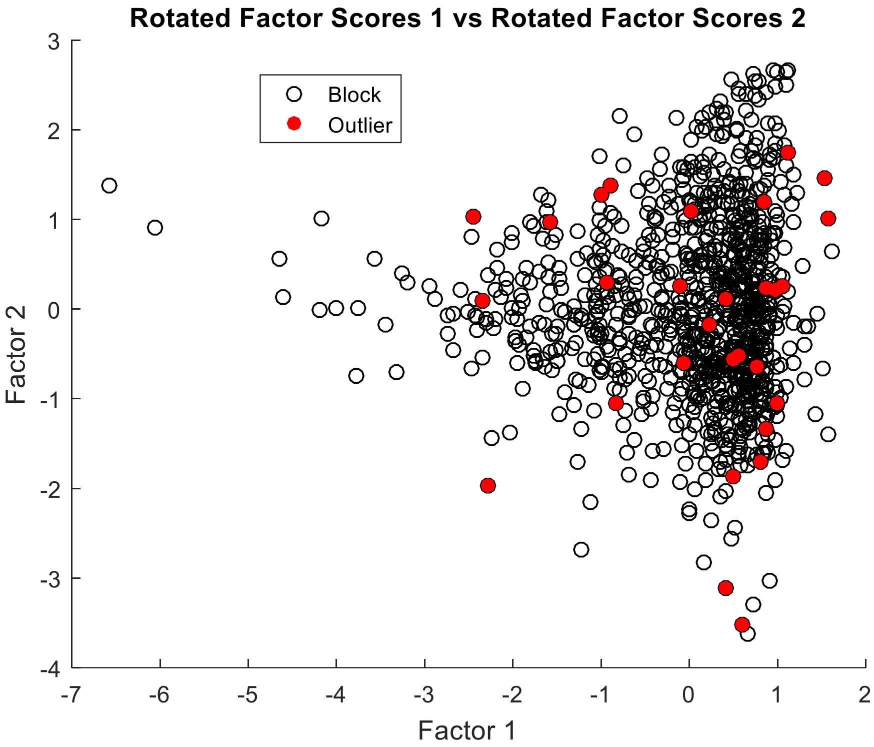


Figure 15: 2D Scatter Plot of Rotated Factors (Updated Data Set)

To compare results across methodology, red dots are plotted corresponding to observations classified as anomalous by the Chi-Square Q-Q plot function. The behavior of these observations within the factor score plot does not reveal any interesting patterns or behavior. Introducing a third factor into the scatter plot, any new emergent patterns or behavior are observed.

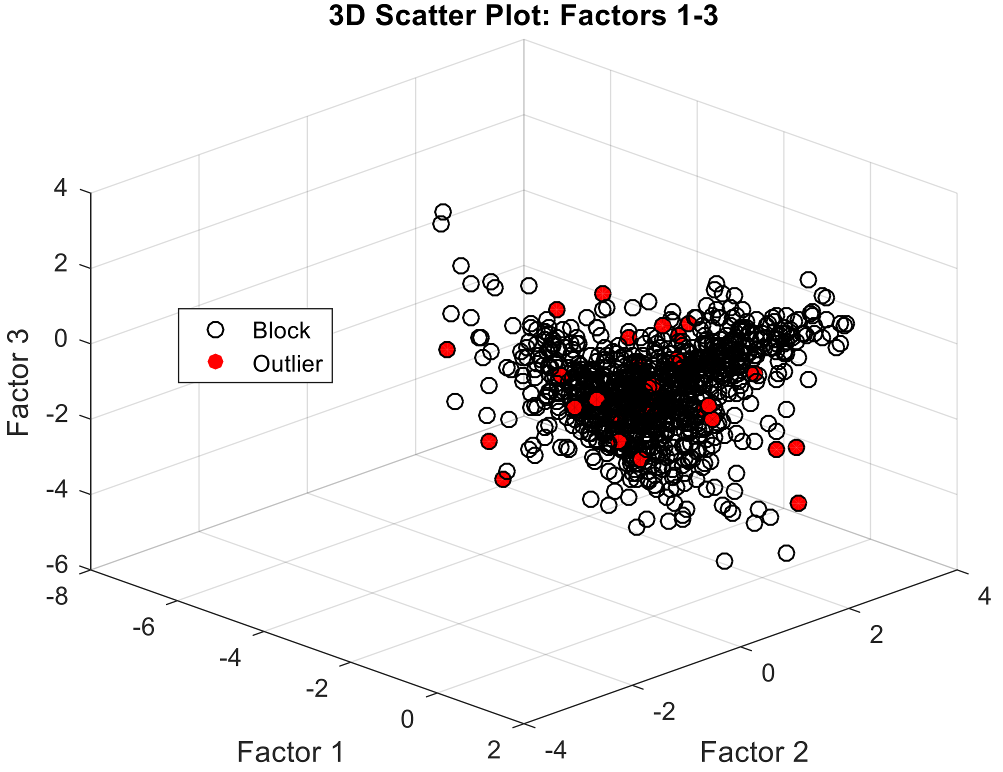


Figure 16: 3D Scatter Plot of Rotated Factors (Updated Data Set)

The inclusion of the first three factors within the data plot does not impact previous result. The observations associated with classified outliers do not hold any distinguishing characteristics, they are very evenly dispersed through the data and do not appear to be out of the ordinary. The fact that there is no discernable pattern within the plot is unsurprising given the severe departure from multivariate normality observed in the updated dataset. When viewing the factor score plots for the original data however, different observations are made.

Comparing with the original dataset, slight differences are noted. As previously observed, the original dataset adheres to the assumption of multivariate normality much better than the updated data. The full battery of factor analysis techniques are implemented as described from the beginning of section 4.3, with interest in the final factor plots. After rotating the factors and finding an optimal factor solution, the resultant 2D factor plot is revealed.

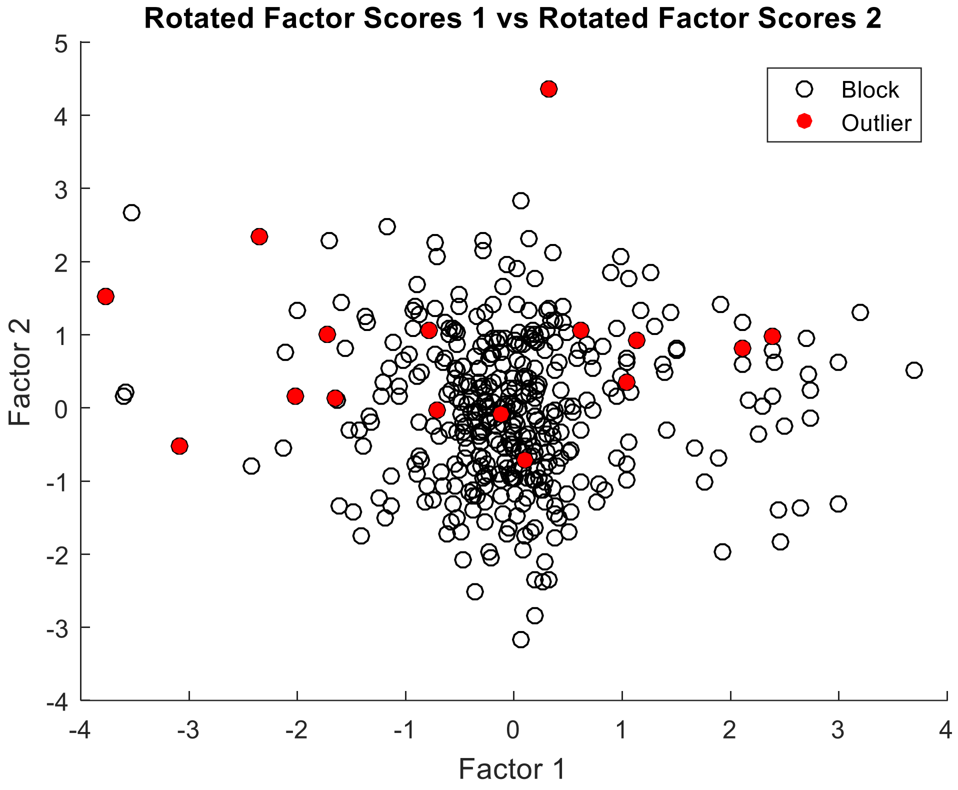


Figure 17: 2D Scatter Plot of Rotated Factors (Original Data Set)

Again, red scatter points indicate where the function classified a vector block as an outlier, and while the difference is not overwhelmingly apparent, there does appear to be a slight pattern within this plot.   Compared with the factor plot in figure 14, there are very few outlier data points present in the central cloud of data. Outliers tend to be concentrated on the perimeter of the scatter cloud at a much greater quantity.   Introducing the third Factor and replotting reveals a consistent phenomenon.

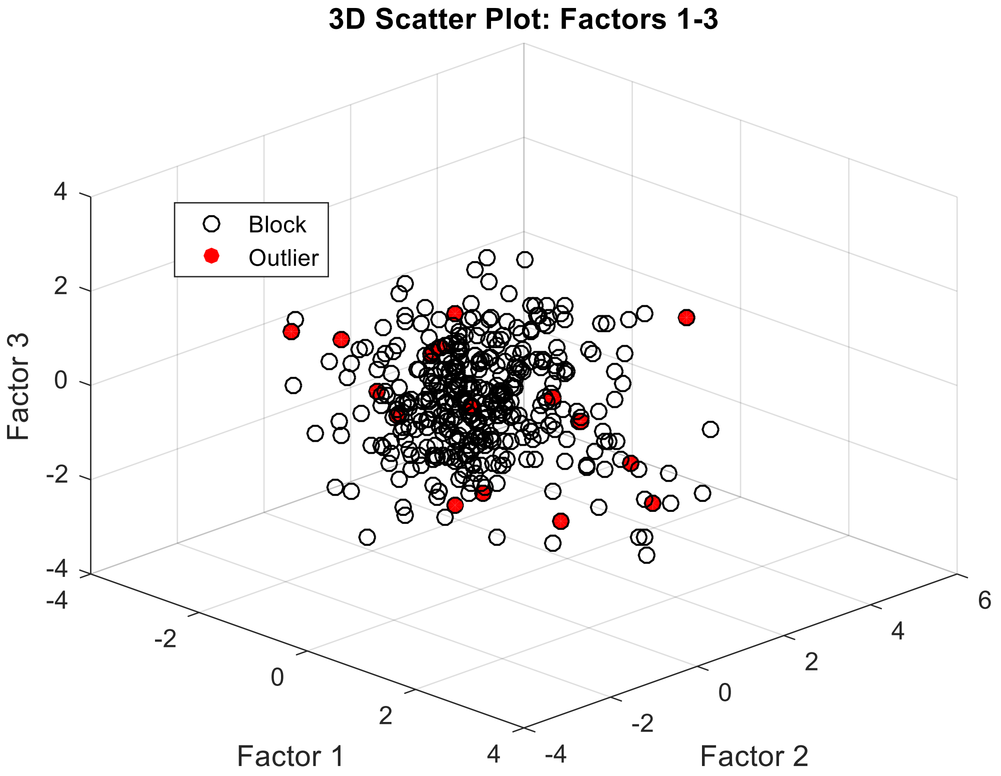


Figure 18: 3D Scatter Plot of Rotated Factors (Original Data Set)

The three-dimensional plot confirms the initial observation, as anomalous data points are heavily focused on the perimeter of the data cloud. There is nothing definitive that can be said about these factor plots and what they yield. There does appear to be more of an outlier pattern formed with the original data, however, that statement is somewhat subjective and inconclusive. At this point, the formation of slight pattern in the original dataset is apparent, where the updated dataset reveals no distinctive patterns at all. These results possibly stem from the differences in the underlying data structures, in which a dataset adhering to multivariate normality will yield conclusive or meaningful phenomenon given that they exist. Based on the factor analysis alone, it is not possible to conclude whether anomalous behavior is being observed, or if meaning is simply being ascribed where there is none. To form a more conclusive opinion on the reliability of the classification, formal multivariate normality testing of the datasets is used.

4.4 Formal Test for Multivariate Normality

After the Chi-Square function executes, a reduced data frame is left from which anomalous observations were removed, and with a separate data frame of outliers. The results of formal multivariate normality testing on the reduced data frame will allow assessment of the original assumptions regarding the two separate data sets. Operating under the assumption that the reduced dataset is not multivariate normal, Mardia’s MVN test and subsequently the Henze-Zirkler MVN test are used.

```
  Mardia's Multivariate Normality Test
---------------------------------------
  data : CurrentVector

  g1p: 692.8973
  chi.skew : 112133.9
  p.value.skew : 0

  g2p: 2867.23
  z.kurtosis: 154.5618
  p.value.kurt: 0

  chi.small.skew: 112495.1
  p.value.small : 0

  Result: Data are not multivariate normal.
---------------------------------------
```

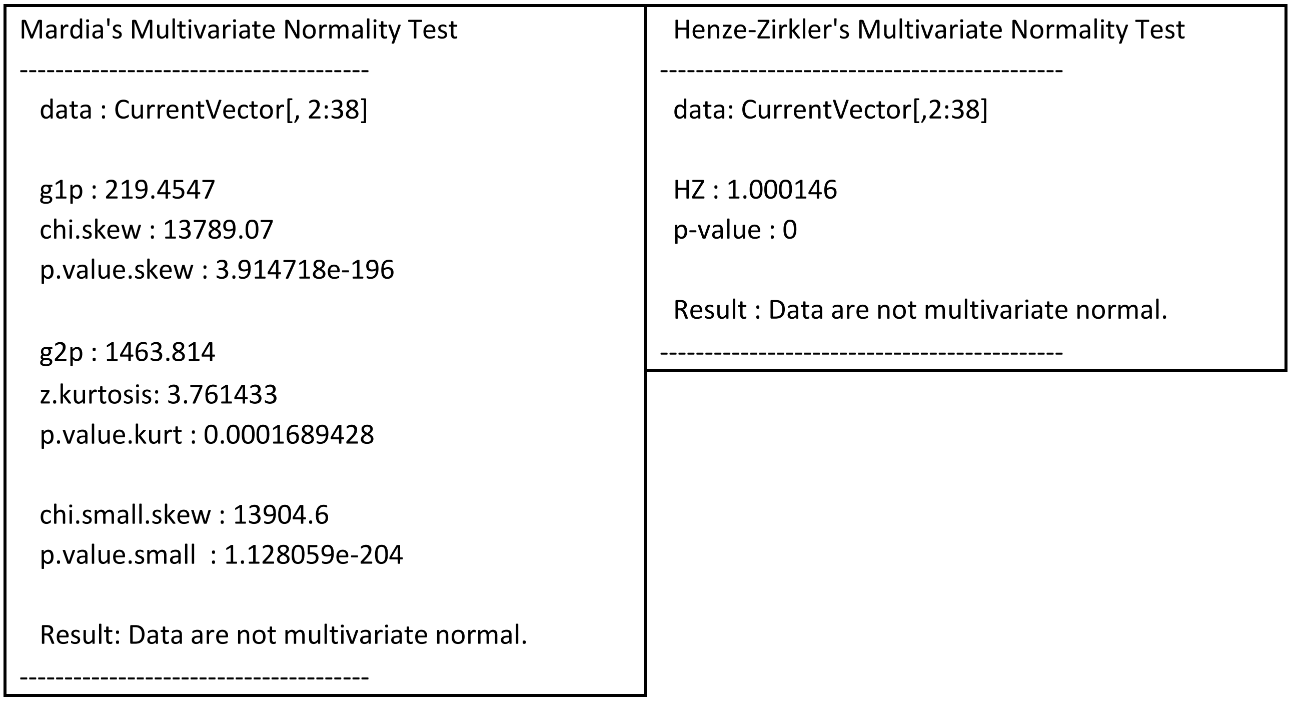
```
  Henze-Zirkler's Multivariate Normality Test
---------------------------------------------
  data : CurrentVector

  HZ : 1.028627
  p-value: 0

  Result: Data are not multivariate normal.
---------------------------------------------
```

Figure 19: MVN Test Results (Updated Data Set)

The results above confirm the initial assessments based on the behavior of the Chi-Square Q-Q plot. It can now be stated with more certainty that the updated dataset is not multivariate normal. This is a good indication that the Chi-Square Q-Q plots are generating accurate visual cues into underlying data structure, as the plot depicted in figure 11 also demonstrates a severe departure from the ideal model. This same formal test is conducted again against the original dataset, and yields unexpected, yet important results.



```
Mardia's Multivariate Normality Test
---------------------------------------
   data : CurrentVector[, 2:38]

   g1p : 219.4547
   chi.skew : 13789.07
   p.value.skew : 3.914718e-196

   g2p : 1463.814
   z.kurtosis: 3.761433
   p.value.kurt : 0.0001689428

   chi.small.skew : 13904.6
   p.value.small  : 1.128059e-204

   Result: Data are not multivariate normal.
---------------------------------------
```

```
 Henze-Zirkler's Multivariate Normality Test
---------------------------------------------
 data: CurrentVector[,2:38]

 HZ : 1.000146
 p-value : 0

 Result : Data are not multivariate normal.
---------------------------------------------
```

Figure 20: MVN Test Results (Original Data Set)

Although the reduced form of the Chi-Square Q-Q plot seen in figure 8 appears to closely match the multivariate normal model, formal testing reveals that the dataset is in fact not multivariate normal. Looking at Mardia's test results gives some insight as to why the dataset fails to satisfy MVN assumptions. The most significant factors contributing to the test results comes from skewness in the dataset. Looking to the original plots, the scale is updated to reveal a more accurate depiction.

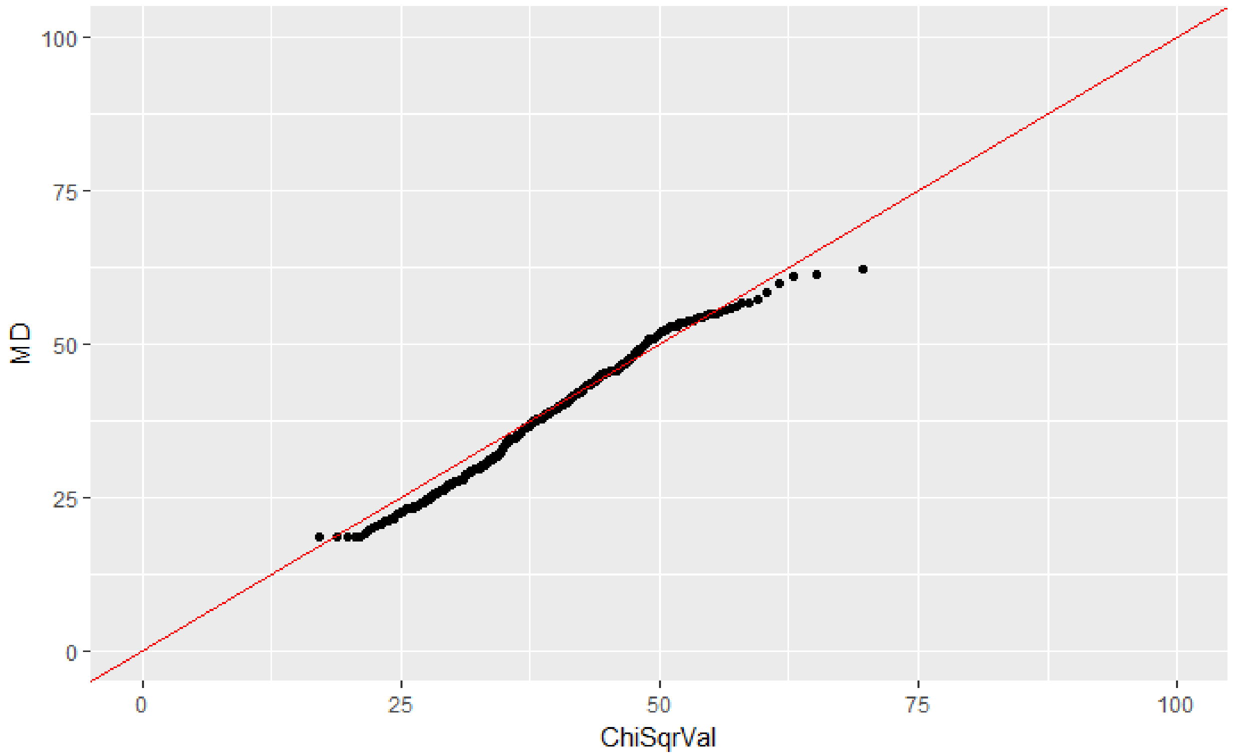


Figure 21: Rescaled Chi-Square Q-Q Plot (Original Data Set)

The `ggplot2` [26] package used to generate the plots will adjust the scale of the plot automatically to a best fit, forcing the scale from 0-100, a plot is obtained that, while fitting well along the model line, only inhabits a small range of the 0-100 scale. In the implementation of the Chi-Square Q-Q plot, it will be important to ensure this scale is used as the standard for plotting, lest it misleadingly lead to a false assumption of multivariate normality.

4.5 Simulation of Multivariate Normality

Since multivariate normal data is not being used, the `MASS` [24] package is implemented to simulate a multivariate normal dataset. Using the variance and column means from a sample dataset built into RStudio®, a multivariate normal matrix consisting of 1000 observations and 11 features is simulated. Finding the MD and Chi-Square distribution vectors, a Chi-Square Q-Q plot is constructed.

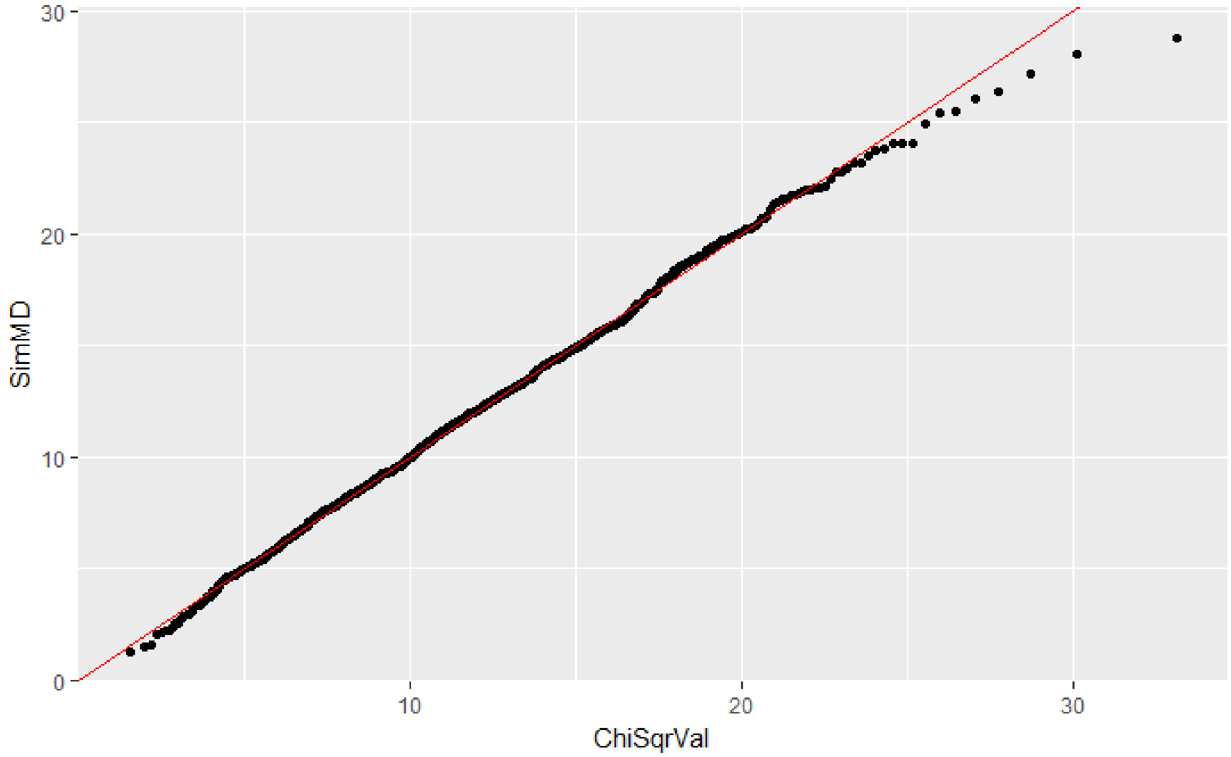
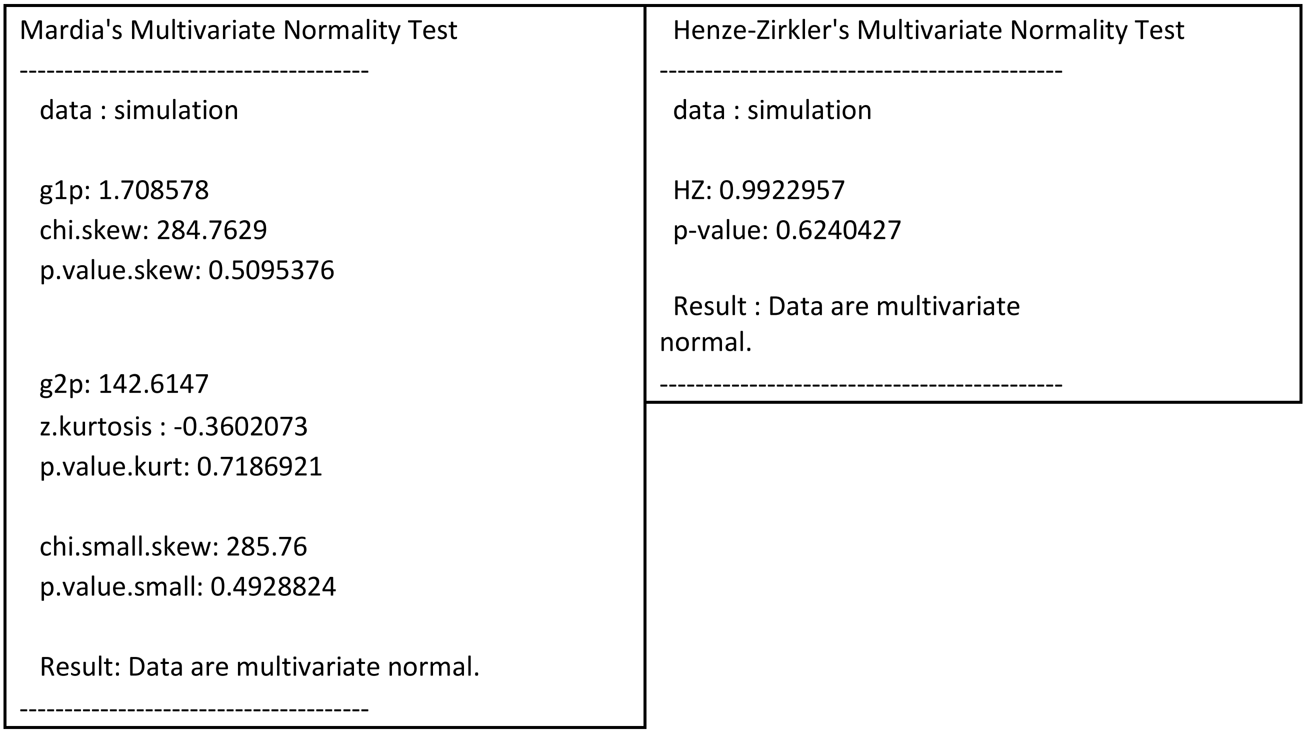


Figure 22: Chi-Square Plot Q-Q (Simulated Data Set)

This graph shows what a multivariate normal dataset would look like on the Chi-Square Q-Q plot. There is a very tight adherence to the perfectly multivariate normal model, and no issues with the plotting scale. There is a consistent 45-degree plot pattern originating from a point close to origin. Calculating the standard error of the estimate for this dataset, a value of .022 is found, which is much smaller than the best error value achieved of 1.62. Formal testing allows for further validation of the results of the Chi-Square Q-Q Plot and error estimate by formally testing the simulated data set for multivariate normality.



```
Mardia's Multivariate Normality Test
---------------------------------------
   data : simulation

   g1p: 1.708578
   chi.skew: 284.7629
   p.value.skew: 0.5095376


   g2p: 142.6147
   z.kurtosis : -0.3602073
   p.value.kurt: 0.7186921

   chi.small.skew: 285.76
   p.value.small: 0.4928824

   Result: Data are multivariate normal.
---------------------------------------
```

```
  Henze-Zirkler's Multivariate Normality Test
---------------------------------------------
  data : simulation

  HZ: 0.9922957
  p-value: 0.6240427

  Result : Data are multivariate
normal.
---------------------------------------------
```

Figure 23: MVN Test Results (Simulated Data Set)

The formal testing validates the result of the Chi-Square Q-Q plot and to a lesser extent, the standard error of the estimate. There is confidence in the conclusion that a close adherence to the model is a strong indication of a multivariate dataset, and that the plot serves as a reliable visual indicator for the structure of the underlying data set. The issues that have presented themselves in this research do not reveal flaws with the techniques, but rather, unreliable data.

4.6 Updated Histogram Matrix

Although results indicate that the data is not suited for this type of multivariate application, it can still be demonstrated what the proposed histogram matrix update would look like. Using outliers classified via the Chi-Square Q-Q plot as depicted in Table 5, the histogram matrix is reconstructed,

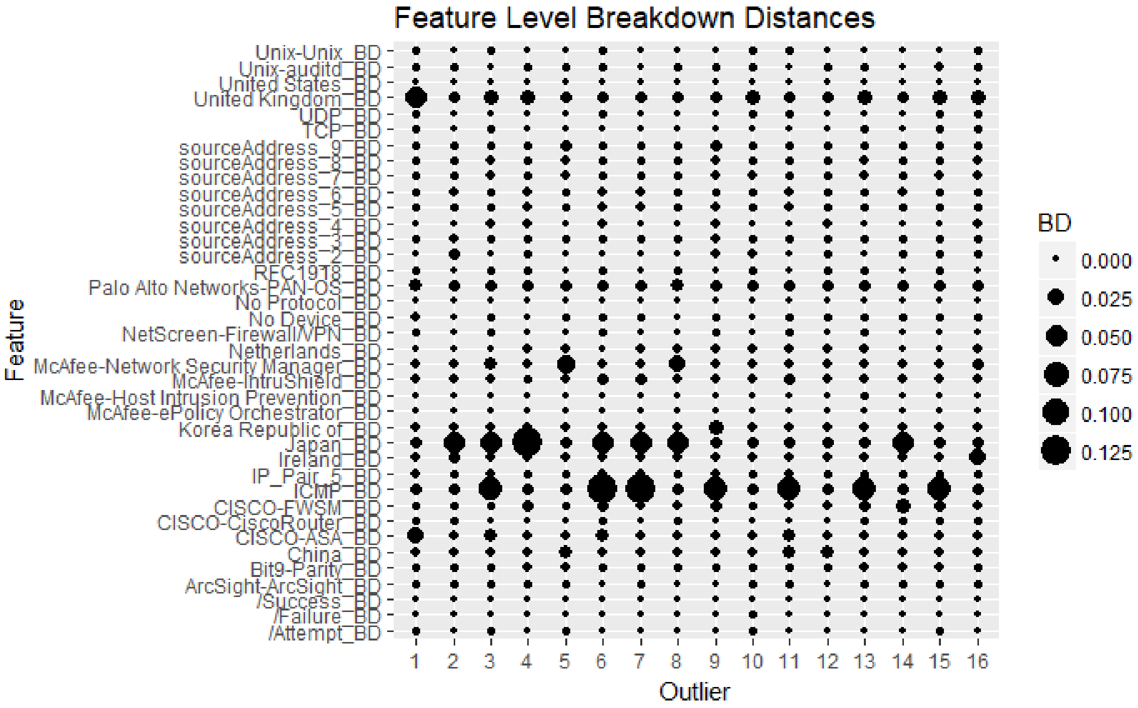


Figure 24: Updated Histogram Plot

This plot in Figure 24 represents an improvement on the original as depicted in Figure 4. The MD color gradient has been eliminated as it represents extraneous information and did not convey it in a concise or objective manner. Due to anomaly classification via the Chi-Square Q-

Q plot, it is known that there were 16 outliers classified, where outlier 1 on the histogram corresponds to the first outlier classified. This simplified histogram allows the user to focus simply on the breakdown distance measure. The user can also reference Table 5 if more information on an outlier is needed. The table will provide the time range during which the anomaly occurred, and the associated block. Eliminating a large quantity of information provides a more concise histogram, absent an ambiguous and distracting color gradient.

V. Conclusion

5.1 Take-Away

The core take-away from this research is that it is possible to apply powerful multivariate analytic solutions to the cyber anomaly detection problem, but the reliability of the results varies with the data being used. With the introduction of a new raw dataset, different variables on which to base analysis can be selected. This introduction of new variables revealed weaknesses in the multivariate analytic based approaches to outlier detection as a drastic non-compliance with normality assumptions was observed. When violating the assumptions of multivariate normality, uncertainty as to the reliability of results is introduced. One benefit of using the Chi-Square plot is in the visual indication given to whether the data are multivariate normal or not. When multivariate normality is not achieved, it does not mean that the techniques demonstrated here are rendered useless or irrelevant, it simply means that they must be applied with discretion; and perhaps validated with alternative multivariate analytic techniques. Despite formal testing revealing that the original dataset is not multivariate normal, plotting the classified outliers within the factor score plot reveals interesting patterns consistent with outlier behavior. Finally, the proposed classification criterion in this body of work eliminates ambiguity associated with which observations should be classified as outliers. Rather than taking a random sample of observations, a rank order list is generated. Outlier rank classification in the updated histogram generated is much more intuitive because of this and does not rely on an ambiguous color gradient.

5.1.1 Contributions

The primary improvement upon this research was the implementation of an outlier classification criteria via a Chi-Square Q-Q plot. Iteratively updating the covariance matrices, as observations are eliminated, it is possible to track how closely the data structure matches the ideal mode. The Q-Q plot functionality makes two major contributions, first, there now exists a definitive criterion for outlier classification. Before this initiative, there was no formal methodology for classification, rather, a user would simply look at the first twenty observations with the highest MD and rank them based on a continuous color gradient. The Chi-Square Q-Q plot classification allows identification of an appropriate number of observations for the dataset, with the only user defined parameter being the maximum percentage of data allowed for classification.

The second major contribution lies in the structure of the Q-Q plot itself. While useful for defining classification parameters, it also offers a strong visual indication as to whether a dataset is multivariate normal or not simply by observing how closely data points adhere to the ideal model line plotted in red. This means that a user can tell without conducting any formal testing, if the results of analysis are reliable or not. This represents a massive asset to an untrained cyber analysis who may not have the educational background to understand the importance of underlying assumptions behind many multivariate techniques. This functionality provides not only a defined classification methodology, but a built-in test for multivariate normality.

5.1.2 Data Considerations

In the results, it was observed that formal multivariate normality testing reveled determinations of abnormality for two separate sets of data based on the IDS/IPS logs. When considering what type of data is ideal for multivariate application, there is a tendency to favor data that is continuous over categorical, or descriptive data. In the raw data files, the primary type of data observed is categorical: event classes, IP addresses, and port numbers are just a few examples of data that cannot be handled on a continuous scale. The introduction of the state vector mechanic transforms these variables into continuous counts, however, this comes at the cost of expanding the dataset massively.

If there are for example, 2000 categorical levels in a feature, then invoking the tabulated state vector functionality, would force the data frame to expand by 2000 features to apply the count mechanic to every single categorical level. Avoiding an unreasonably large data frame, limits are set on how many categorical levels can be converted into features, however, with the updated dataset, this still expands the data frame quite rapidly. Before adjustment for multicollinearity, the tabulated state vector for the updated dataset boasts 215 features for 1000 observations. So far, it has not been possible to establish a set of features which demonstrate multivariate normality, and part of this may be because the raw data sets simply contain too many categorical features that are not neatly converted into a format compatible with multivariate analysis

Within the raw data, there may exist a combination of multivariate normal features, however, identifying what those might be is not productive if the features are irrelevant to the analyst. Throughout this research, there was some disconnect as to which features should be selected for analysis. Sponsors originally advised the use of the same features from Gutierrez's research as depicted in table 1, however, upon exploration of the newly provided raw data, there is an absence of many of those original features. It is possible to test for and select data which fits well in a multivariate analytic based model, however, this would not ensure meaningful results. It is imperative to keep sponsor feedback within the loop of research especially regarding feature selection. They should be the final authority over which features are retained for analysis and which are ignored. By maintaining this feedback loop, the most useful analytic tools for outlier detection both in capability and relevance can be designed.

5.2 Future Research Considerations

Perhaps one of the most important contributions that can be added to this research is in validation of results. While searching for anomalies via a wide array of analytic techniques, there is currently no way of validating outlier classifications. In an ideal research scenario, there would have dataset with several observations that are known to be anomalous. Fabricating datasets with outliers to showcase the efficacy of analytic techniques is possible, however, actual datasets relevant to sponsors are of much greater interest. In the case that this type of data is made available, a high priority should be placed on validation of methods.

The techniques afforded by multivariate analysis represent only one way in which users might obtain outlier determinations within a data set. Given what is known about the underlying structure of the tabulated state vector data sets, beneficial research would be any that attempts utilizing alternate means for outlier classification. The construction of a tailored neural network is just one example of an alternate technique that may yield appreciable results.

VI. Deliverable

The final result of this research will be in the production of a web-based user application built in the R Shiny environment. This application is being developed to satisfy OPER 782 requirements and will allow users to implement several of the features discussed in this research. The primary focus of this application will be on the execution of the Chi-Square Plot function, and on exportation of classified outliers. The repository in which the code for the application is being held is available on GitHub via the following link: https://github.com/citation891/MCAC. The project title 'MCAC' stands for Multivariate Chi-Square Anomaly Classification. A delivery schedule for the proposed shiny app is available in the repository readme.md file.

Appendix A: Rotated Lambda Factor Loadings

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4						
/Access	-0.80	0.05	-0.23	-0.02	-0.39	-0.02	0.00	-0.07	0.01	0.21
/Host/Res/Mem	-0.11	0.01	-0.04	-0.47	0.01	0.02	0.00	0.01	0.00	0.00
/Network	-0.25	-0.03	-0.06	-0.12	-0.87	-0.07	0.00	0.09	0.00	0.08
NA3	0.02	-0.15	0.08	-0.10	-0.05	-0.08	0.00	0.35	0.01	0.01
/DM/MA/Thrsh/Rising	0.07	-0.28	0.05	-0.07	-0.04	0.08	0.00	-0.05	0.00	0.00
/DM/MA/Val/Current	0.14	-0.72	0.20	-0.05	0.03	-0.14	0.00	-0.11	-0.01	-0.01
/M/A/EPS/PA	-0.04	-0.05	0.03	-0.40	-0.01	-0.12	0.00	0.11	0.00	0.01
/M/A/EPS/PostFilter	-0.12	0.01	-0.02	-0.37	-0.06	-0.10	0.00	0.04	0.00	0.01
/M/A/EPS/Received	-0.01	-0.07	-0.03	-0.48	-0.06	0.03	0.00	0.01	0.00	0.01
/M/A/EPS/ToManager	-0.06	-0.02	-0.06	-0.46	0.03	0.15	0.00	-0.06	0.00	0.01
/M/A/Events/ToManager	0.01	-0.02	-0.01	-0.47	-0.04	-0.01	0.00	-0.06	0.00	0.01
NA4	0.03	-0.13	0.04	-0.17	-0.11	0.12	0.00	0.00	0.00	0.00
BR	0.06	0.05	-0.19	0.01	0.06	0.05	0.00	-0.04	0.00	0.00
CA	-0.17	-0.19	-0.57	0.05	-0.24	0.16	0.00	0.24	0.01	0.01
CH	0.04	0.00	-0.01	-0.03	-0.07	0.11	0.00	0.51	0.00	0.00
CZ	0.28	-0.11	0.07	-0.04	-0.04	0.10	0.00	-0.19	-0.01	0.00
DE	0.02	-0.02	-0.06	-0.33	0.08	-0.19	0.00	0.00	0.00	-0.01
DK	0.04	0.05	-0.22	-0.11	-0.19	-0.45	0.00	0.00	-0.01	0.01
FR	0.03	-0.06	0.05	0.03	0.08	-0.50	0.00	-0.02	0.00	0.00
GB	0.17	0.02	-0.01	-0.14	0.04	-0.07	0.00	-0.11	0.00	0.00
GT	0.12	0.39	0.23	0.10	0.06	0.06	-0.01	0.60	0.02	0.00
HR	0.37	0.62	0.27	0.16	0.11	0.39	-0.03	-0.28	0.07	-0.01
IE	0.04	-0.11	-0.09	0.06	-0.13	-0.03	0.00	-0.01	-0.01	0.01
JP	-0.07	-0.04	-0.29	-0.10	-0.15	-0.37	0.00	-0.08	-0.01	0.01
KR	-0.67	-0.23	-0.02	0.02	-0.32	-0.24	0.00	-0.14	-0.08	0.05
LB	-0.05	-0.01	-0.39	0.06	-0.27	0.13	0.00	-0.06	0.00	-0.03
NL	-0.03	-0.02	-0.12	-0.07	0.02	-0.63	0.00	0.00	0.01	0.00
RU	-0.13	-0.05	-0.03	-0.06	0.01	-0.56	0.00	-0.03	-0.01	0.00
SA	0.03	0.03	-0.05	-0.42	-0.03	-0.15	0.00	0.05	0.00	-0.01
SG	0.00	0.01	0.06	-0.01	-0.05	-0.12	0.00	-0.03	0.01	0.00
US	-0.22	-0.77	-0.15	-0.06	0.08	0.14	-0.01	0.03	-0.08	0.00
NA5	0.07	-0.29	0.04	-0.64	-0.04	-0.04	0.00	-0.05	0.00	-0.02
rule:105	-0.25	-0.01	-0.04	-0.14	-0.87	-0.07	0.00	0.07	0.02	-0.08
Attack: Suspicious Source	-0.12	-0.05	0.01	-0.12	0.04	-0.18	0.00	0.10	0.00	0.00
Traffic: Dark Add. Space	-0.89	0.06	-0.19	0.05	-0.09	-0.05	0.00	-0.10	0.04	-0.14
COUNT_EVENT	0.12	-0.54	0.13	0.00	-0.02	-0.11	0.00	0.00	0.00	0.00
IP_DST_2	-0.58	-0.11	-0.03	-0.63	-0.02	0.01	0.00	-0.06	0.01	-0.06
IP_DST_3	-0.82	0.00	-0.03	-0.23	-0.01	-0.05	0.00	-0.11	-0.01	0.01
IP_DST_4	-0.69	0.04	-0.01	-0.16	-0.03	0.02	0.00	-0.03	0.01	-0.01
IP_SRC_1	0.33	0.76	0.32	0.17	0.10	0.18	0.00	-0.15	-0.22	0.00
IP_SRC_2	0.09	0.51	-0.14	0.15	0.08	0.08	0.00	0.60	-0.03	-0.02
IP_SRC_3	-0.14	0.16	-0.70	0.04	0.10	0.01	0.00	0.25	0.00	0.00
IP_SRC_4	-0.17	0.00	-0.80	0.00	0.07	-0.10	0.00	0.06	0.00	0.01
IP_SRC_5	-0.14	-0.01	-0.80	-0.12	-0.04	-0.19	0.00	-0.08	0.00	0.01
IP_SRC_6	-0.03	0.02	-0.76	-0.18	-0.11	-0.16	0.00	-0.16	-0.01	0.01
IP_SRC_7	0.01	0.00	-0.63	-0.18	-0.13	-0.18	0.00	-0.19	0.00	0.01

VII. Bibliography

[1] M. Ahmed, A. Naser Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *Journal of Network and Computer Applications*, vol. 60. pp. 19–31, 2016.

[2] K. W. Bauer, “OPER 685 (Applied Multivariate Analysis) Course Notes.” Air Force Institute of Technology, Wright Patterson AFB OH, 2016.

[3] T. J. Bihl, W. A. Young II, and G. R. Weckman, “Defining, Understanding, and Addressing Big Data,” *International Journal of Business Analytics*, vol. 3, no. 2, pp. 1–32, 2016.

[4] B. Boehmke and R. Gutierrez, “anomalyDetection: Implementation of Augmented Network Log Anomaly Detection Procedures," *The R Journal*, vol. 9, no. 2, pp. 354–365, 2017.

[5] A. Costello and J. Osborne, *Best practices in exploratory factor analysis*, 1st ed, Createspace Publishing, 2004.

[6] W. Dillon and M. Goldstein, *Multivariate Analysis: Methods and Applications*, 1st ed. Canada: John Wiley & Sons, Inc., 1984.

[7] P. Filzmoser and M. Gschwandtner, “‘mvoutlier’: Multivariate outlier detection based on robust methods.” CRAN, p. 48, 2015.

[8] A. Frei and M. Rennhard, “Histogram Matrix: Log file visualization for anomaly detection,” *ARES 2008 - 3rd International Conference on Availability, Security, and Reliability, Proceedings*, no. April 2008, pp. 610–617, 2008.

[9] R. G. Garrett, “The chi-square plot: a tool for multivariate outlier recognition,” *Journal of Geochemical Exploration*, vol. 32, no. 1–3, pp. 319–341, 1989.

[10] H. D. Gibson, S. G. Hall, and G. S. Tavlas, “A suggestion for constructing a large time-varying conditional covariance matrix,” *Economics Letters*, vol. 156, pp. 110–113, 2017.

[11] R. Gnanadesikan, *Methods for statistical data analysis of multivariate observations*, vol. 321. New York: John Wiley and Sons, 2011.

[12] R. J. Gutierrez, “a Tabulated Vector Approach for Log-Based Anomaly Detection,” Air Force Institute of Technology, 2017.

[13] N. Henze and B. Zirkler, “A class of invariant consistent tests for multivariate normality,” *Communications in Statistics - Theory and Methods*, vol. 19, no. 10, pp. 3595–3617, 1990.

[14] Kaiser H, “Analysis of factorial simplicity,” *Psychometrika*, vol. 39, no. 1, pp. 31–36, 1974.

[15] T. Kollo, “Multivariate skewness and kurtosis measures with an application in ICA,” *Journal of Multivariate Analysis,* vol. 99, no. 10, pp. 2328–2338, 2008.

[16] S. Korkmaz, D. Goksuluk, and G. Zararsiz, “MVN: An R package for assessing multivariate normality,” *The R Journal*, vol. 6, no. 2013, pp. 151–162, 2014.

[17] D. M. Lane, “Standard error of the estimate,” in *Online Statistics Education: A Multimedia Course of Study,* Rice University, 2014, pp. 8–10.

[18] P. C. Mahalanobis, “On the Generalized Distance in Statistics,” *National Institute of Sciences,* vol. 2, pp. 49–55, 1936.

[19] K. V. Mardia, “Measures of multivariate skewness and kurtosis with applications” *Biometrika,* vol. 57, no. 3, pp. 519–530, 1970.

[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis,* 5th ed. Hoboken: John Wiley & Sons, Inc., 2006.

[21] R. J. Rummel, *Applied factor analysis*. Evanston, IL: Northwestern University Press, 1970.

[22] K. Scarfone and P. Mell, “Guide to Intrusion Detection and Prevention Systems (IDPS) Recommendations of the National Institute of Standards and Technology,” *NIST Special Publication*, vol. 800–94, p. 127, 2007.

[23] A. Urbelis, “WannaCrypt ransomware attack should make us wanna cry,” *CNN* 17-May-2017.

[24] W. N. Venables and B. D. Ripley, “Modern Applied Statistics With S,” *Technometrics,* vol. 45, no. 1. Springer, New York, pp. 111–111, 2003.

[25] H. Wickham, “readxl: Read Excel files,” *R package version 0.1*, 2016.

[26] H. Wickham, “ggplot2 Elegant Graphics for Data Analysis,” *Media*, vol. 35, July. Springer-Verlag New York, p. 211, 2009.

[27] H. Wickham, “tidyverse: Easily Install and Load ‘Tidyverse’ Packages,” *R package version 1.0.0*, 2016.

[28] H. Wickham and R. Francois, “dplyr: A Grammar of Data Manipulation,” *R Package Version 0.7.4,* 2016.

[29] Y. Xie, “knitr: A General-Purpose Tool for Dynamic Report Generation in R,” *The R Journal* vol. 8, no. 1, pp. 1–12, 2012.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 22-03-2018	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From - To)* Sep 2016 – Mar 2018

4. TITLE AND SUBTITLE Outlier Classification Criterion for Multivariate Cyber Anomaly Detection	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Trigo, Alexander M, 1Lt	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865	**8. PERFORMING ORGANIZATION REPORT** AFIT-ENS-MS-18-M-166
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Army Cyber Command Fort Gordan GA 30905 LTC Cade Saie Cade.m.saie.mil@mail.mil	**10. SPONSOR/MONITOR'S ACRONYM(S)** ASC **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** N/A

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A. Approved for Public Release; Distribution Unlimited.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Every day, intrusion detection systems catalogue millions of unsupervised data entries. This represents a "big data" problem for research sponsors within the Department of Defense. In a first response to this issue, raw data capture was transformed into usable vectors and an array of multivariate techniques implemented to detect potential outliers…

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	73	**19b. TELEPHONE NUMBER** *(include area code)* 937-271-4242 bradleyboehmke@gmail.com